

AD-A119 110

AIR FORCE INST OF TECH WRIGHT-PATTERSON AFB OH  
TWO-DIMENSIONAL NUMERICAL SIMULATION OF SEMICONDUCTOR DEVICES.(U)  
MAY 82 C H PRICE  
AFIT/CI/NR/82-24D

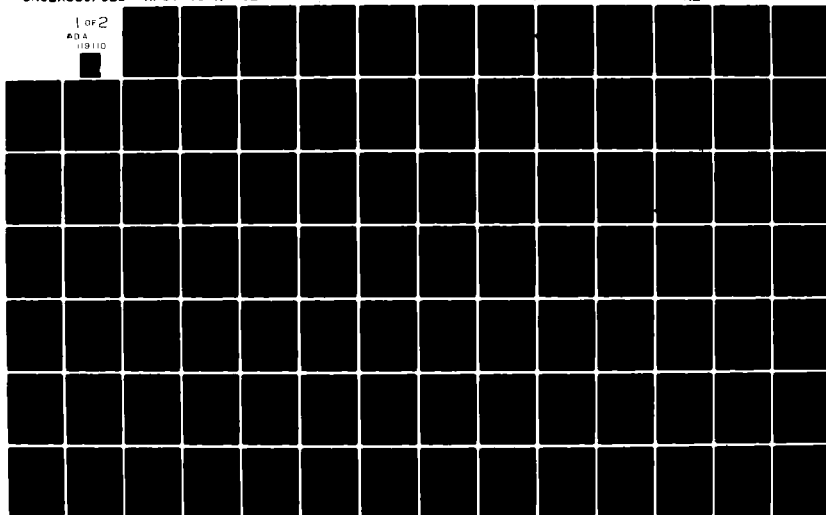
F/6 9/1

UNCLASSIFIED

NL

1 of 2

AD A  
119 110



UNCLASS

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER AFIT/CI/NR/82-24D	2. GOVT ACCESSION NO. AD-211911C	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) Two-Dimensional Numerical Simulation of Semi-conductor Devices	5. TYPE OF REPORT & PERIOD COVERED THESIS/DISSERTATION	
7. AUTHOR(s) Craig H. Price	6. PERFORMING ORG. REPORT NUMBER	
9. PERFORMING ORGANIZATION NAME AND ADDRESS AFIT STUDENT AT: Stanford University	8. CONTRACT OR GRANT NUMBER(s)	
11. CONTROLLING OFFICE NAME AND ADDRESS AFIT/NR WPAFB OH 45433	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS	
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)	12. REPORT DATE May 1982	
	13. NUMBER OF PAGES 133	
	15. SECURITY CLASS. (of this report) UNCLASS	
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE	
6. DISTRIBUTION STATEMENT (of this Report) APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED		
17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)		
18. SUPPLEMENTARY NOTES APPROVED FOR PUBLIC RELEASE: IAW AFR 190-17 30 AUG 1992		
19. KEY WORDS (Continue on reverse side if necessary and identify by block number)		
20. ABSTRACT (Continue on reverse side if necessary and identify by block number) ATTACHED		

DTIC

ELECTE

SEP 9 1982

H

Lynn E. Wolaver

Dean for Research and  
Professional Development  
AFIT, Wright-Patterson AFB OH

FORM 1473 JAN 73

EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASS

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

30

## ABSTRACT

Two-dimensional numerical simulation is a necessary tool for modern semiconductor device design. Analytical models and judicious application of one-dimensional simulation cannot accurately represent the highly two-dimensional impurity profiles and structures of VLSI devices. Moreover, the allowable device structures and bias conditions of existing two-dimensional simulation programs are too restrictive to provide the necessary design information.

A two-dimensional numerical simulation program, PISCES, has been written in order to study various aspects of device simulation. The program uses vectorized  $LU$  decomposition to alternately solve Poisson's equation and the electron current continuity equation (Gummels method). The program is extremely flexible and useful in evaluating two-dimensional simulation concerns such as grid allocation, boundary conditions, convergence characteristics and physical models.

The discretization<sup>↑</sup> grid is analyzed in comparisons of rectangular and triangular grids and in the allocation of grid points within critical regions of the device. A triangular grid achieved by distorting a rectangular grid is advocated as a reasonable compromise between the flexibility of general triangular grids and the regularity and matrix solution method compatibility of rectangular grids. A finite difference discretization of Poisson's equation and the current continuity equation on a triangular grid is presented.

A variety of methods for reducing the simulation time are explored. The nested dissection grid renumbering scheme is shown to provide a significant storage and operation count reduction for larger grids with a slight penalty in vector operation efficiency. Techniques for accelerating the convergence of

the alternating method are presented which together reduce solution times by a factor of four for devices biased above threshold. These methods involve computation of an improved initial guess, elimination of excessive solutions of Poisson's equation, overrelaxation of the potential updates and reduction of the Poisson linearization term. Even with these improvements, however, simulations above threshold still require about four times as long as subthreshold simulations. This slow convergence appears to be correlated with slow oscillations of the first harmonic in spatial frequency of the surface potential in the inverted channel between source and drain.

Two application examples demonstrate the utility of the PISCES program and two-dimensional numerical simulation in general. Simulation of an implanted channel MOSFET reveals a 50 fold increase in punchthrough current with a 12% increase in source drain junction depth. Field dependent mobility is investigated with the implementation of a distance-from-the-surface mobility model.



Accession For	
NTIS GTR&I	<input checked="checked" type="checkbox"/>
DTIC TAB	<input type="checkbox"/>
Unannounced	<input type="checkbox"/>
Justification	
By _____	
Distribution/	
Availability Codes	
Dist	Avail and/or Special
A	

## AFIT RESEARCH ASSESSMENT

The purpose of this questionnaire is to ascertain the value and/or contribution of research accomplished by students or faculty of the Air Force Institute of Technology (AFIT). It would be greatly appreciated if you would complete the following questionnaire and return it to:

AFIT/NR  
Wright-Patterson AFB OH 45433

RESEARCH TITLE: Two-Dimensional Numerical Simulation of Semi-conductor Devices

AUTHOR: Craig H. Price

## RESEARCH ASSESSMENT QUESTIONS:

1. Did this research contribute to a current Air Force project?  
☐ a. YES ☐ b. NO
2. Do you believe this research topic is significant enough that it would have been researched (or contracted) by your organization or another agency if AFIT had not?  
☐ a. YES ☐ b. NO
3. The benefits of AFIT research can often be expressed by the equivalent value that your agency achieved/received by virtue of AFIT performing the research. Can you estimate what this research would have cost if it had been accomplished under contract or if it had been done in-house in terms of manpower and/or dollars?  
☐ a. MAN-YEARS \_\_\_\_\_ ☐ b. \$ \_\_\_\_\_
4. Often it is not possible to attach equivalent dollar values to research, although the results of the research may, in fact, be important. Whether or not you were able to establish an equivalent value for this research (3. above), what is your estimate of its significance?  
☐ a. HIGHLY SIGNIFICANT ☐ b. SIGNIFICANT ☐ c. SLIGHTLY SIGNIFICANT ☐ d. OF NO SIGNIFICANCE
5. AFIT welcomes any further comments you may have on the above questions, or any additional details concerning the current application, future potential, or other value of this research. Please use the bottom part of this questionnaire for your statement(s).

NAME \_\_\_\_\_ GRADE \_\_\_\_\_ POSITION \_\_\_\_\_

ORGANIZATION \_\_\_\_\_ LOCATION \_\_\_\_\_

STATEMENT(s):

FOLD DOWN ON OUTSIDE - SEAL WITH TAPE

AFIT/NR  
WRIGHT-PATTERSON AFB OH 45433  

---

OFFICIAL BUSINESS  
PENALTY FOR PRIVATE USE. \$300



NO POSTAGE  
NECESSARY  
IF MAILED  
IN THE  
UNITED STATES

**BUSINESS REPLY MAIL**

FIRST CLASS PERMIT NO. 73236 WASHINGTON D.C.

POSTAGE WILL BE PAID BY ADDRESSEE

AFIT/ DAA  
Wright-Patterson AFB OH 45433



FOLD IN

TWO-DIMENSIONAL NUMERICAL SIMULATION  
OF SEMICONDUCTOR DEVICES

A DISSERTATION  
SUBMITTED TO THE DEPARTMENT OF ELECTRICAL ENGINEERING  
AND THE COMMITTEE ON GRADUATE STUDIES  
OF STANFORD UNIVERSITY  
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS  
FOR THE DEGREE OF  
DOCTOR OF PHILOSOPHY

By  
Craig H. Price

May 1982


© Copyright 1982

by

Craig H. Price



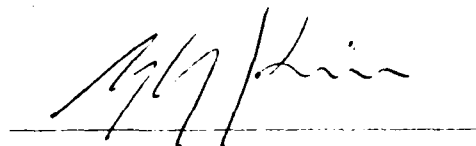
I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

  
(Principal Adviser)

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.



I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.



Approved for the University Committee  
on Graduate Studies:

\_\_\_\_\_  
Dean of Graduate Studies

## ACKNOWLEDGEMENTS

I would like to express my sincere appreciation to my advisor, Professor Robert W. Dutton, for his assistance and support throughout the course of this work. Thanks are also due to Professor James D. Plummer and Professor Gordon S. Kino for their critical reading of the manuscript.

I am extremely grateful to the members of the Integrated Circuits Laboratory for their suggestions and assistance, especially to James Greenfield for his valuable ideas and insight and to Dr. Soo-Young Oh for providing direction in the early stages of this work. Special thanks also to Stephen Hansen and Andrew Lanza for outstanding programming and computer system support. I am also grateful to Col. James Carpenter for his cooperation, support and patience during the many months of manuscript preparation.

I would also like to thank my parents, family and friends for their faith and encouragement. But most of all, I wish to thank my wife Rebecca for her enduring support throughout these past five years. And to my daughter, Meredith, my apologies for the hours which we could not share.

This work was supported by the United States Air Force through the Air Force Institute of Technology / Civilian Institutions program and by the Defense Advanced Research Projects Agency (contract MDA 903-79-C-0257).

## TABLE OF CONTENTS

1. INTRODUCTION . . . . .	1
1.1. Perspective . . . . .	1
1.2. History . . . . .	9
1.3. PISCES . . . . .	12
1.4. Overview . . . . .	14
2. GRID . . . . .	17
2.1. Grid Types . . . . .	17
2.2. Grid Density Criteria . . . . .	23
2.3. Boundary Sensitivities . . . . .	27
2.4. Summary . . . . .	35
3. DISCRETIZATION . . . . .	36
3.1. Poisson's Equation . . . . .	39
3.1.1. Area Allocation . . . . .	45
3.1.2. Obtuse Triangles . . . . .	47
3.2. Continuity Equation . . . . .	50
3.2.1. Electron Transport Equation . . . . .	50
3.2.2. Electron Continuity Equation . . . . .	52
3.2.3. Obtuse Triangles . . . . .	55
3.3. Summary . . . . .	58

4. SOLUTION TECHNIQUES . . . . .	60
4.1. Matrix Equation Solution Methods . . . . .	60
4.1.1. Poisson's Equation . . . . .	61
4.1.2. Continuity Equation . . . . .	68
4.1.3. Renumbering Algorithms . . . . .	69
4.2. Solution of Coupled Equations . . . . .	76
4.2.1. Solution Methods . . . . .	77
4.2.2. Convergence Acceleration for the Alternating Method . . . . .	79
4.2.2.1. Projection of the Initial Guess . . . . .	81
4.2.2.2. Single Poisson's Equation Iteration . . . . .	84
4.2.2.3. Overrelaxation . . . . .	6
4.2.2.4. Linearization Term Reduction . . . . .	
4.2.3. Convergence Rate Sensitivity to Bias Conditions . . . . .	
4.3. Summary . . . . .	
5. APPLICATIONS . . . . .	99
5.1. MOSFET Punchthrough . . . . .	99
5.2. Strong Inversion Mobility . . . . .	103
5.3. Summary . . . . .	110
6. CONCLUSION . . . . .	111
6.1. Summary . . . . .	111
6.2. Recommendations . . . . .	114
Appendix A. PISCES DEMONSTRATION EXAMPLE . . . . .	117
REFERENCES . . . . .	133

## LIST OF TABLES

4.1	Nested Dissection Efficiencies . . . . .	74
4.2	Nested Dissection Efficiencies versus Grid Size . . . . .	75
4.3	Nested Dissection Effect on Vector Arithmetic . . . . .	76
4.4	Subthreshold Convergence . . . . .	80
4.5	Linear Region Convergence . . . . .	82
4.6	Number of Matrix Solutions to Convergence . . . . .	92
4.7	Approximate Solution Times on IIP-1000F . . . . .	97

## LIST OF FIGURES

1.1	A comparison of short- and long-channel MOSFET equipotential lines. . . . .	3
1.2	Charge sharing model of Yau. . . . .	4
1.3	Modern two-dimensional device structures: (a) static induction transistor, (b) taper-isolated dynamic-gain RAM cell. . . . .	6
1.4	The parallel paths of simulation versus actual wafer processing. Feedback is provided at each level by the comparison of measured or simulated results on profiles or operating characteristics against the desired values. . . . .	8
1.5	Examples of PISCES graphical output (unretouched) showing equipotential lines for (a) MOSFET and (b) MESFET. . . . .	15
2.1	Various rectangular and triangular grids. . . . .	19
2.2	Overlaying grid on non-rectangular structures. . . . .	21
2.3	Potential along the insulator-semiconductor interface for a device in punchthrough ( $V_G = -.1V$ , $V_{BG} = 3V$ , $V_S = 0V$ , $V_D = 4V$ ). . . . .	25
2.4	PISCES grid for an IGFET. . . . .	27
2.5	Boundary condition sensitivities in the linear region; $V_G = 2.4V$ , $V_{BG} = V_S = 0V$ , and $V_D = .01V$ . . . . .	28
2.6	Lateral impurity profile at the semiconductor surface. . . . .	29
2.7	Current sensitivity to boundary conditions in the subthreshold and linear regions for the structures of Figure 2.5. . . . .	31
2.8	Boundary condition sensitivities in the saturation region; $V_G = 2V$ , $V_{BG} = V_S = 0V$ , and $V_D = 5V$ . . . . .	32
2.9	Current sensitivity to boundary conditions in the saturation region for the structures of Figure 2.8. . . . .	33
3.1	Sample grid with five triangles. The triangles are labeled $t_1$ to $t_5$ , $A_i$ is the area associated with the central node, and $l_i$ is the boundary of that area. . . . .	40

3.2	Labeling convention for a triangle. The distance between nodes is $d$ , the length (height) of the boundary segments is $h$ , and $\hat{u}$ is a unit vector. . . . .	42
3.3	Equivalence of flux boundary $s$ to flux boundaries $h_j$ plus $h_k$ . . . .	46
3.4	Various boundaries with equivalent flux conservation characteristics. . . . .	46
3.5	Dependence of area weighting on triangle orientation using the centroid method. The area allocation in case (b) is twice that of case (a). . . . .	47
3.6	Coupling coefficients for Poisson's equation on an obtuse triangle. . . . .	48
3.7	Area allocation with a mix of acute and obtuse triangles. . . . .	50
3.8	Labeling convention for one-dimensional current transport. . . . .	52
3.9	Coupling coefficient derivation for the continuity equation on acute and obtuse triangles. . . . .	56
4.1	Structure of coefficient matrix for: (a) rectangular grid, (b) rectangular based triangular grid. The dashed lines represent partially filled matrix diagonals resulting from the choice of the / or \ rectangle diagonals for subdivision into triangles. . . . .	70
4.2	Nested dissection numbering scheme for a 7 by 7 grid. Part (a) shows the partitioning and part (b) the actual numbering of the grid. . . . .	71
4.3	Coefficient matrix structure for the nested dissection of Figure 4.2 on a rectangular grid. . . . .	72
4.4	Coefficient matrix structure for the nested dissection of Figure 4.2 on a rectangular based triangular grid. . . . .	73
4.5	Algorithm flow of simultaneous and alternating methods for solution of the coupled equations. . . . .	78
4.6	Algorithm flow of the single Poisson acceleration scheme showing elimination of the inner loop. . . . .	85

4.7	Potential and quasi-Fermi potential convergence of a node in the channel region of a MOSFET biased in saturation. . . . .	87
4.8	Finer detail of the convergence shown in Figure 4.7 including node electron concentration convergence and total drain current convergence. . . . .	87
4.9	Drain current convergence acceleration using overrelaxation. . . .	88
4.10	Drain current convergence acceleration using reduction of the linearizing term. . . . .	91
4.11	Subthreshold and linear region simulation results at $V_{DS} = .01V$ for .2V increments of $V_G$ . . . . .	91
4.12	Saturation region simulation results at $V_G = 2V$ for .5V increments of $V_{DS}$ . . . . .	94
4.13	Surface potential error at each iteration (of the first 18) for a MOSFET in saturation. The channel extends from approximately .75 $\mu m$ to 1.75 $\mu m$ . . . . .	96
5.1	Equipotential contour plot of N-channel MOSFET with implanted channel. . . . .	100
5.2	MOSFET punchthrough characteristics using PISCES, GEMINI, and CADDET. The current path in the .4 $\mu m$ junction device is at the semiconductor surface for biases below 7V and in the bulk for higher biases. The .45 $\mu m$ junction device results are from PISCES only. . . . .	101
5.3	Rolloff of drain current with gate bias showing effect of distance-from-the-surface mobility model. The mobility is in $cm^2/V\cdot s$ and $y$ is in $\text{\AA}$ . . . . .	106
5.4	Definition and relationship between effective mobility $\mu_{eff}$ and field effect mobility $\mu_{FE}$ . . . . .	107
5.5	Comparison of simulated and measured values of $\mu_{eff}$ and $\mu_{FE}$ with varying gate voltage. . . . .	108
5.6	Comparison of simulated and measured values of $\mu_{max}$ sensitivity to substrate doping. . . . .	109



A.1	Sample PISCES input card deck for mesh generation and device structure definition. . . . .	118
A.2	Sample PISCES input card deck for specifying device material characteristics and obtaining simulation solutions. . . . .	119
A.3	Plot of the mesh generated by the PISCES example. . . . .	126
A.4	Equipotential contours generated by the example at $V_G = 2V$ and $V_{DS} = 2V$ . . . . .	132
A.5	Quasi-Fermi potential contours generated by the example at $V_G = 2V$ and $V_{DS} = 2V$ . . . . .	132

## LIST OF SYMBOLS

$A_i$	area allocated to node $i$
$A_{im}$	portion of area $A_i$ in triangle $t_m$
$C_0$	gate capacitance per unit area
$\vec{D}$	electric flux density
$D_n, D_p$	electron (hole) diffusion constant
$d_{jm}, d_{km}$	distance between node $i$ and node $j$ ( $k$ ) in triangle $t_m$
$E$	matrix of error values from approximate decompositions
$\vec{E}$	electric field
$E_g$	energy gap
$\vec{E}_m$	electric field in triangle $t_m$
$E_x, E_y$	magnitude of electric field in $x$ ( $y$ ) direction
$g_d$	drain conductance
$g_m$	transconductance
$G_n, G_p$	electron (hole) generation rate
$h_{jm}, h_{km}$	length of $\vec{l}_{imjm}$ ( $\vec{l}_{imkm}$ )
$I_D$	drain current
$\vec{J}_n, \vec{J}_p$	electron (hole) current density
$J_{njm}, J_{nkm}$	electron current density along edge $ij$ ( $ik$ ) in triangle $t_m$
$J_{nx}, J_{ny}$	electron current density in the $x$ ( $y$ ) direction
$k$	Boltzmann's constant
$k$	Fourier harmonic number
$L$	electrical channel length
$L$	lower triangular matrix
$l$	Fourier solution region width
$l_i$	boundary of area allocated to node $i$

$l_{im}$	portion of boundary $l_i$ in triangle $t_m$
$\vec{l}_{imjm}, \vec{l}_{imkm}$	length weighted vector normal to that portion of $l_i$ in triangle $t_m$ perpendicular to triangle side $ij$ ( $ik$ )
$M$	total number of triangles containing node $i$
$M$	difference equation coefficient matrix
$m$	number of rows in a grid
$N$	net ionized impurity concentration
$N$	total number of nodes in a grid
$N_i$	net ionized impurity charge density in area $A_i$
$N_{inv}$	inversion layer carrier concentration per unit area
$n$	free electron concentration
$n$	number of columns in a grid
$n_i$	intrinsic carrier concentration
$n_i, n_j$	free electron concentration at node $i$ ( $j$ )
$n_{jm}, n_{km}$	free electron concentration at node $j$ ( $k$ ) of triangle $t_m$
$p$	free hole concentration
$p_i$	free hole concentration at node $i$
$Q_I$	inversion layer charge per unit area
$Q_B$	bulk depletion charge per unit area
$q$	electronic unit of charge
$R_n, R_p$	electron (hole) recombination rate
$T$	temperature
$U$	upper triangular matrix
$U_n, U_p$	net electron (hole) recombination rate
$\vec{u}_{jm}, \vec{u}_{km}$	unit vector in direction of $\vec{l}_{ir,jm}$ ( $\vec{l}_{imkm}$ )
$\vec{u}_x, \vec{u}_y$	unit vector in the $x$ ( $y$ ) direction
$V_B$	barrier height

$V_{DS}$	drain to source voltage
$V_G$	gate voltage
$V_T$	thermal voltage
$W$	electrical channel width
$\vec{x}$	solution vector in matrix equation
$\vec{y}$	constant vector in matrix equation
$\alpha$	extrapolation factor for projection of an initial guess of the simulation solution
$\alpha$	term in the energy gap temperature dependence
$\beta$	term in the energy gap temperature dependence
$\epsilon$	permittivity
$\epsilon_m$	permittivity within triangle $t_m$
$\epsilon_s$	permittivity within the semiconductor
$\mu_b$	bulk mobility
$\mu_{eff}$	effective channel mobility
$\mu_{FE}$	field effect mobility
$\mu_{max}$	maximum channel mobility
$\mu_n, \mu_p$	electron (hole) mobility
$\mu_s$	mobility exactly at the semiconductor surface
$\rho$	net charge concentration
$\rho_i$	net charge density in $A_i$
$\rho^k$	$k^{th}$ Fourier harmonic of charge density
$\sigma$	characteristic length of depth dependent mobility variation
$\phi_n, \phi_p$	electron (hole) quasi-Fermi level
$\phi_{ni}, \phi_{pi}$	electron (hole) quasi-Fermi level at node $i$
$\psi$	electrostatic potential
$\psi_i$	electrostatic potential at node $i$

$\psi_{jm}, \psi_{km}$       electrostatic potential at node  $j$  ( $k$ ) in triangle  $t_m$   
 $\psi^k$                  $k^{th}$  Fourier harmonic of potential

## Chapter 1

### INTRODUCTION

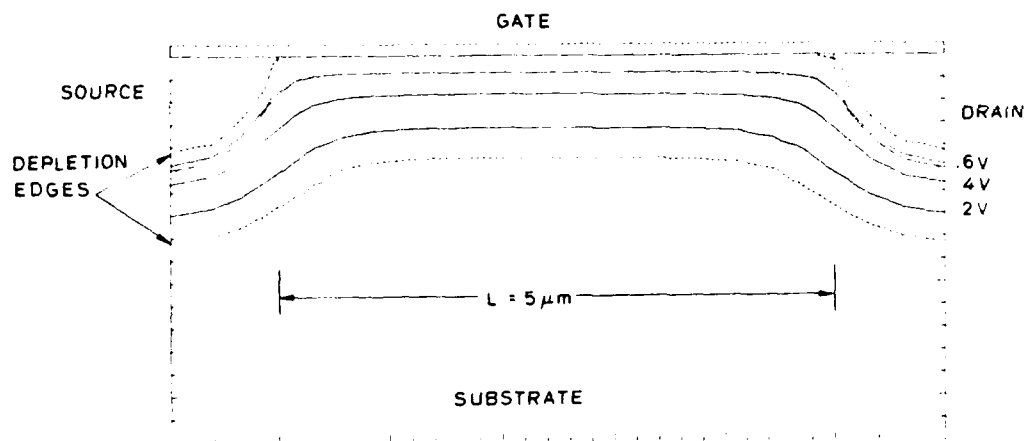
As device geometries shrink in the pursuit of Very Large Scale Integration (VLSI), two-dimensional numerical simulation of devices gains in importance. One-dimensional approximations which are valid for low fields and large lateral dimensions with respect to vertical dimensions no longer apply. Extensions to the one-dimensional theory can be useful but cannot accurately account for the highly two-dimensional structure of modern devices.

#### *1.1 Perspective*

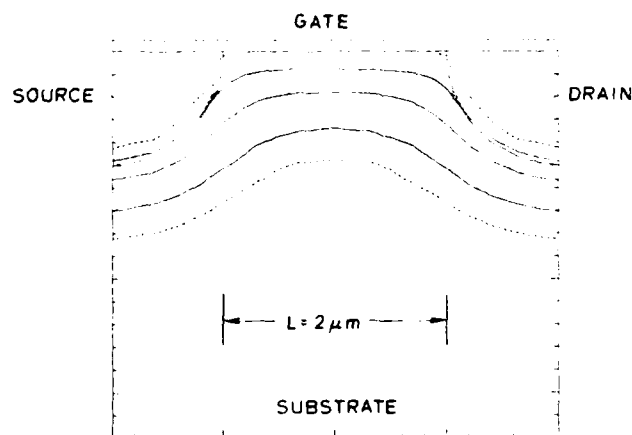
In 1970 gate lengths of 7 microns were typical for production LSI circuits [1.1]. By 1980 the gate length had been reduced to 2 microns and current contracts for the US Government's Very High Speed Integrated Circuits (VHSIC) program [1.2] call for production of 0.5 micron devices by 1985. The constant reduction of device dimensions seen over the last 20 years is expected to continue to at least 1990. As the device lateral dimensions have come down it has become necessary to alter other structural and operational parameters in order to maintain desirable operating characteristics. Device scaling theory [1.3] describes how to optimally adjust device vertical dimensions, voltage levels, and doping concentrations as a function of lateral dimensions to minimize the short-channel effects caused by strong two-dimensional fields. Practical considerations, however, including fabrication constraints and logic level noise immunity have caused designers to sub-optimally scale these parameters thus retaining some short-channel behavior.

Figure 1.1 shows a comparison of the equipotential contours of two metal-oxide-semiconductor field effect transistors (MOSFET's) which are identical except for their gate length. For Figure 1.1a, the channel length (metallurgical junction spacing) is 5 microns while for Figure 1.1b it is 2 microns. Note that in both cases, as the equipotential lines near the surface curve to follow the junction boundaries, they pull away from the surface indicating higher surface potentials near the junctions and an increase in control over these potentials by the source and drain and a reduction in control by the gate. For the long-channel device of (a), these edge effects are a small percentage of the total channel length and thus have limited influence on device characteristics. In the short-channel device, however, the edge effects extend throughout a large percentage of the channel length and have a strong influence on the device characteristics. One result of this is a lower threshold voltage than that predicted by theory since the surface potential is higher for a given gate bias. Another result is an increased sensitivity of the output current to the drain bias (drain conductance) due to the control exerted by the drain on the surface potential in the channel.

*Such effects are clearly the result of the two-dimensional structure, and attempts have been made to model these effects analytically. Some success has been obtained by Yau [1.4] and extended by others [1.5] to model short-channel effects by using a charge sharing theory. The basic features of this model are shown in Figure 1.2. The charge controlled by the gate is assumed to be contained within the trapezoidal region immediately beneath the gate with the remaining charge controlled by the source and drain respectively. Further, the junctions and depletion edge boundaries are assumed to be cylindrical.*



(a)



(b)

Fig. 1.1. A comparison of short- and long-channel MOSFET equipotential lines.

This model roughly approximates the threshold shifts seen as the channel length is reduced in devices fabricated using conventional long-channel techniques; however, as device structures are optimized for short-channel performance, the assumptions made (e.g. cylindrical junctions) no longer apply and the model is invalidated. Thus with analytical modeling, if all of



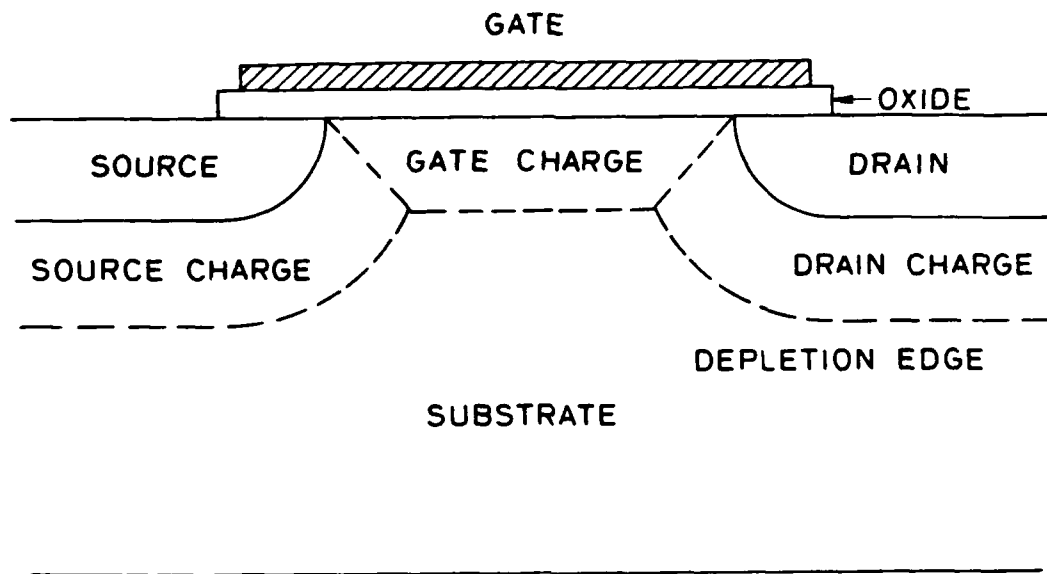


Fig. 1.2. Charge sharing model of Yau.

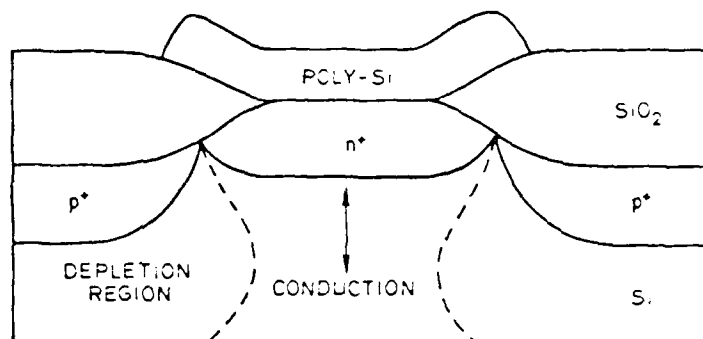
the assumptions made are valid, then the proper results will be obtained; however, if the assumptions break down then the results will be invalid. The principal advantage of numerical simulation of semiconductor devices is that few assumptions need to be made. If there are design errors which result in excessively large electric fields or alternate conduction paths, for example, the simulation will accurately reveal the existence of these effects even though they had not been expected. The principal disadvantages of numerical simulation are that there are no simple equations to describe device behavior, it is not always obvious how to interpret the results, and a significant amount of computation is required.

Of course, attempts can and will be made to model new structures. Analytical models provide insight into the interplay of device parameters which can be useful to device designers and users; however, the modeling job

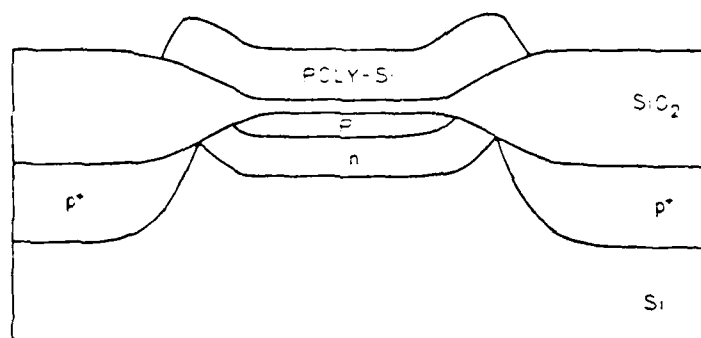
becomes increasingly difficult and the assumptions become more restrictive as device structures shrink in size and become more complex. In fact, numerical simulation is inevitably used in arriving at or verifying analytical models.

As an example of the complexity of modern device structures, Figure 1.3 shows two exotic field effect devices: a static induction transistor [1.6] and a taper-isolated dynamic-gain RAM cell [1.7]. The operating characteristics of these devices result primarily from the two-dimensional nature of their impurity profiles and electric fields. The creation of analytical models for these devices, applicable throughout their entire operating regime, would clearly be a difficult task.

For some applications, (e.g. circuit simulation programs) empirical relations suffice for explaining device behavior. This form of modeling suffers from several drawbacks. First, the model parameters often have no physical basis and may only be extracted from measured data. *Second, since each region of device operation requires a different empirical relation, it is difficult if not impossible to match the device characteristics in the transition from one region to another.* Going a step further, one recent circuit simulation program uses tables instead of closed form expressions for device characteristics [1.8]. Much physical basis is lost in this method except in the determination of how to parameterize the tables. In either case, closed-form expressions or tables, numerical simulation can be used to generate the required device characteristics. Numerical simulation also provides a significant opportunity to study the effects of device technology variables on circuit performance. Furthermore, the use of process simulation to generate the device profiles provides the opportunity to directly study the link between fabrication steps and circuit performance.



(a)



(b)

Fig. 1.3. Modern two dimensional device structures: (a) static induction transistor, (b) taper-isolated dynamic-gain RAM cell.

Two-dimensional numerical simulation of semiconductor devices is a natural and necessary extension of current trends in computer aided design (CAD) as a link between process simulators such as SUPREM [1.9] and circuit simulators such as SPICE [1.10]. Figure 1.4 shows how device simulation fits into a total simulation philosophy. The actual fabrication of devices is preceded by the processing of test structures and measurement of the one-dimensional impurity profiles. Differences from desired profile values are

fed back into the process specifications and the sequence is repeated until satisfactory agreement is obtained. The device lot wafers are then processed and electrical measurements are made from which two-dimensional profile information may be inferred. Electrical measurements are also made to determine device and circuit characteristics. Feedback is provided at each level to allow optimization of the process.

The principal savings in cost and time for simulation versus actual fabrication and testing of devices comes in process simulation. Typically, one simulation of a process using SUPREM would take only a few minutes, costing tens-of-dollars on a mainframe computer. Actual fabrication would typically take several weeks and cost thousands of dollars. Obviously, the savings resulting from the use of simulation are substantial. The situation is reversed somewhat for device and circuit simulation versus device and circuit measurements. Device and circuit electrical measurements are relatively quick and inexpensive tasks while the cost of each simulation is roughly comparable to that for process simulation. Even so, device simulation offers a tremendous advantage over device measurement since one need not fabricate the device first. Moreover, simulation provides a detailed two-dimensional view of the physics which device measurements cannot provide.

Another advantage of simulation lies in the fact that it is often difficult to accurately measure two-dimensional device structures [1.11]. When attempting to analyze or improve device performance, one would like to measure device profiles. In cases where accurate measurements are not possible, simulation allows the engineer the opportunity to manipulate the simulated profiles and observe the effects on device characteristics, thus inferring the actual profile shapes [1.12].

Finally, simulation provides a much greater insight into device behavior

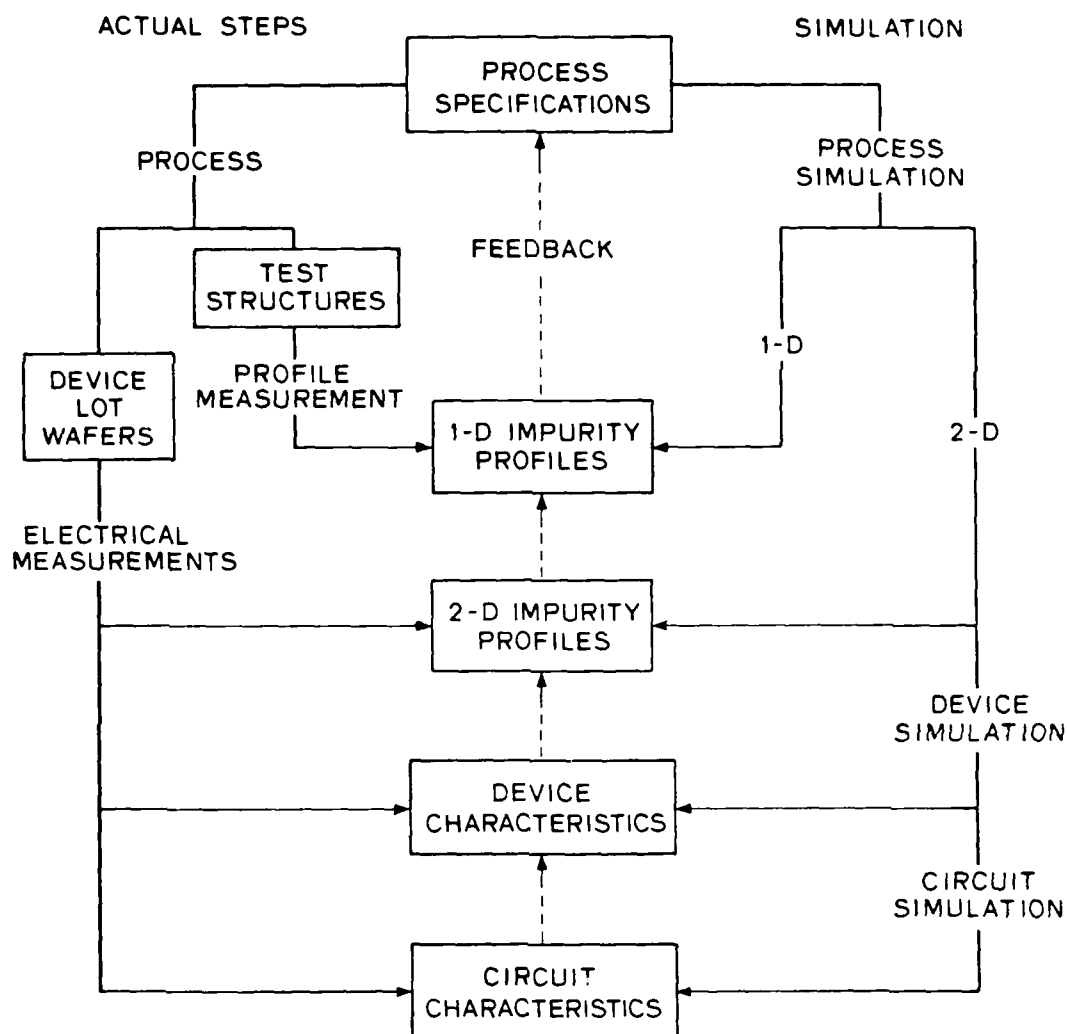


Fig. 1.4. The parallel paths of simulation versus actual wafer processing. Feedback is provided at each level by the comparison of measured or simulated results on profiles or operating characteristics against the desired values.

than can be obtained in any other way. In the case of unwanted punchthrough currents in MOSFETs [1.13], for example, one can both determine terminal

conditions of punchthrough and spatially identify the point of punchthrough. Further, a proposed solution to the problem may be simulated and its success or failure determined.

## 1.2 History

The history of numerical simulation of semiconductor devices begins with the work of Gummel [1.14] in 1964 on the one-dimensional steady-state analysis of bipolar transistors. His method provided a two-carrier solution meaning that the continuity and transport equations were satisfied for both holes and electrons. The principal contribution of his early work was the use of a sequential iteration scheme for obtaining a consistent solution to the three sets of equations: Poisson's equation, and the hole and electron continuity equations. In his method, one first solves Poisson's equation followed by the electron and hole continuity equations in succession. The cycle is then repeated. The solution of each equation individually in this manner requires much less work than solving all three simultaneously. Generally, however, the convergence is not as fast as the quadratic convergence which can be obtained for a simultaneous solution [1.15]. Nevertheless, the amount of work saved in each iteration generally compensates for the slower convergence. These points will be examined in greater detail in Chapter 4.

De Mari analyzed the p-n junction in one dimension in 1968 [1.16] and enhanced the numerical analysis capabilities to include transient conditions [1.17]. In 1969 Scharfetter and Gummel published their work on the transient analysis of a Read diode oscillator [1.18]. This paper provided the second major advance in numerical algorithms with the introduction of a carrier transport equation discretization scheme which allowed larger grid spacing

and thus fewer variables. This method will be described further in Chapter 3. The advent of two-dimensional simulation in 1969 lowered interest in one-dimensional simulation, thus slowing its development and focusing its application to bipolar devices. The most significant works to follow emphasized specific device analyses obtained from one-dimensional simulations and the inclusion of higher order physical phenomena such as band gap narrowing and mobility variations [1.19, 1.20]. Selected works in one-dimensional analysis include Gokhale in 1970 [1.21]; Hachtel, *et al.* in 1972 [1.22]; and D'Avanzo in 1979 [1.23].

Two-dimensional simulation appeared in the literature in 1969 with the publication of works by Kennedy and O'Brien [1.24-26] on the simulation of JFETs and by Slotboom [1.27, 1.28] on bipolar transistors. The advent of short channel FETs in this period was the driving factor for two-dimensional simulation and almost all subsequent work was aimed at IGFETs. Reiser introduced two-dimensional transient analysis in 1970 [1.29-34] and Mock presented his stream function formulation of the carrier transport equations in 1973 [1.34].

The next significant development came in 1973 with the first publication of finite element analysis by Hachtel [1.35, 1.36] followed in 1974 by Barnes and Lomax [1.37, 1.38] and Buturla and Cottrell [1.39-41]. All previous work had been based on finite-difference discretization schemes and their associated rectangular grids. The use of finite elements provided improvements over finite-difference discretization with the ability to model non-rectangular structures and a more efficient use of grid. Although rectangular grids do not prohibit the simulation of non-rectangular structures [1.42], the numerical techniques for accomodating these structures had not been implemented. Hence, previous work had been restricted to rectangular structures. These

topics will be discussed further in Chapter 2.

Developments since 1975 have focused on decreasing program size and solution time while increasing the accuracy of the solution in terms of both the numerical algorithms used and the models of the device physics. Another aim has been the development of "friendly" user interfaces for the simulation programs in order to make the simulation technology more available to device designers [1.43]. This progression from use of simulation as a laboratory tool in developing semiconductor device theory to application in production facilities for optimizing device structures is a field still in its infancy. Nonetheless, its application holds great promise in providing the increased productivity needed for VLSI design.

Several programs have recently become widely available. NEMOS, based on the early work of Kennedy [1.24]; CADDET, based on Mock's work [1.44]; and MINIMOS, [1.45] all provide steady-state solutions for essentially rectangular FET structures. TWIST [1.46] and GEMINI [1.42] solve only for the device potentials but are quite useful for simulation of sub-threshold and punchthrough characteristics as well as device breakdown characteristics. TWIST is limited to rectangular geometries while GEMINI allows non-rectangular structures.

Currently, work is in progress on two fronts which will provide valuable aid to the device designer - the development of three-dimensional simulation and the introduction of simplified fast two-dimensional simulation. Several researchers have published preliminary work on three-dimensional simulation [1.47-49]. Most notable of these is that of Buturla and Cottrell with their extension of the two-dimensional simulation program FIELDAY [1.50] to three dimensions. These simulations have shown that there are characteristics of short and narrow semiconductor devices which can be ac-



curately simulated only in three dimensions. These programs, however, must run on large mainframe computers and consume considerable computer time; thus, they are currently of limited practical value to device designers.

At the other end of the spectrum is the simplified two-dimensional program of Oh, SDVICE [1.51]. This program provides a very fast solution to the two-dimensional FET transient problem and uses very little computer memory. It is limited in the device geometries and operating regions which it handles and the accuracy of its solutions require further verification. However, this program solves a particular type of problem extremely efficiently.

### *1.3 PISCES*

A two-dimensional numerical simulation program has been written for the purpose of investigating grid and boundary condition sensitivities, convergence limitations, device physics models, and to compare other numerical simulation programs. PISCES (Poisson and Single-carrier Continuity Equation Solver) solves the Poisson equation and the steady-state electron continuity equation using the alternating method (Gummel's algorithm) on an HP-1000F minicomputer. The program uses a finite difference discretization on an irregular triangular grid and thus easily handles non-planar surfaces and interfaces.

The program was written for use on field effect transistors where a single-carrier solution is sufficient. Field effect transistors are majority carrier devices and very little error is introduced by ignoring the minority carriers except in extreme biasing conditions such as avalanche breakdown. By solving only the electron continuity equation, time is saved since the hole con-

tinuity equation does not have to be solved, and slightly faster convergence is obtained.

The equations solved are:

$$\begin{array}{ll}
 \text{Poisson} & \vec{\nabla} \cdot (\epsilon \vec{\nabla} \psi) = n - p - N \\
 \text{Continuity} & \vec{\nabla} \cdot \vec{J}_n = q(R_n - G_n) \\
 \text{Transport} & \vec{J}_n = -q\mu_n n \vec{\nabla} \psi + qD_n \vec{\nabla} n \\
 \text{Boltzmann} & n = n_i e^{q(\psi - \phi_n)/kT} \\
 & p = n_i e^{q(\phi_p - \psi)/kT}
 \end{array}$$

where  $\epsilon$  is the permittivity,  $\psi$  is the electric potential,  $n$  and  $p$  are the free electron and hole concentrations,  $N$  is the net ionized impurity concentration,  $\vec{J}_n$  is the electron current density,  $R_n$  and  $G_n$  are the electron recombination and generation rates,  $\mu_n$  is the electron mobility,  $D_n$  is the electron diffusion constant,  $n_i$  is the intrinsic carrier concentration,  $\phi_n$  and  $\phi_p$  are the electron and hole quasi-Fermi levels, and  $kT/q$  is the thermal voltage.

The extension of finite difference methods to triangular grids is a novel approach for semiconductor device simulation although it has been applied in other fields [1.52]. Irregular triangular grids possess desirable features for semiconductor simulation including the ability to conform to irregular shapes and the capacity for local grid refinement without inducing excessive grid elsewhere. These same advantages are the driving factors which have resulted in the development of finite element simulators.

The PISCES program also contains user oriented features which increase its utility. Using the input parser and graphics interface routines developed for the GEMINI program, PISCES provides a flexible, friendly user interface for both input and output. Two examples of contour plot output from PISCES are shown in Figure 1.5. Equipotential contours are shown for a

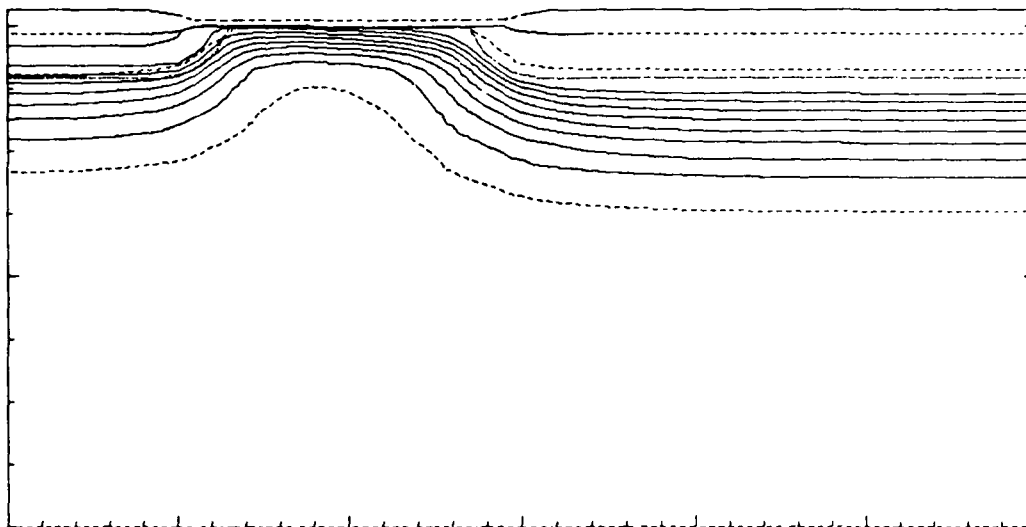
typical short-channel MOSFET and for a MESFET. A complete PISCES simulation example is provided in Appendix A.

#### *1.4 Overview*

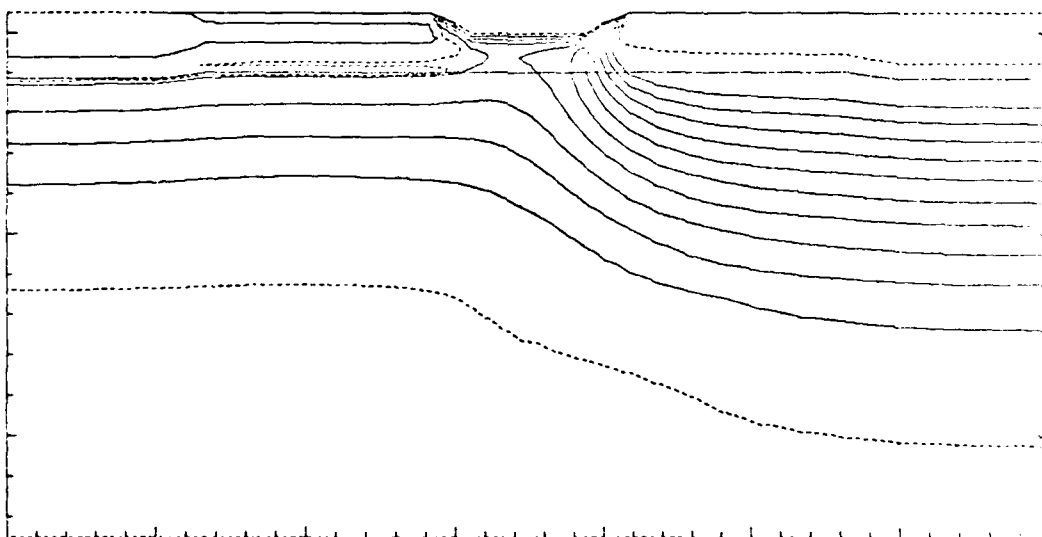
The objective of this work is to aid the development of device simulation by examining various aspects of the problem. A versatile two-dimensional simulation program has been developed in the course of this study. The versatility is achieved through choice of grid and discretization schemes which allow simulation of modern, highly two-dimensional, non-planar semiconductor devices on a minicomputer. The use of a minimum number of grid points and application of several novel convergence acceleration techniques reduces solution times to practical limits.

The effects of grid on two-dimensional numerical simulation of semiconductor devices are discussed in Chapter 2. The various types of grid are presented and their applicability to device structures and solution methods are considered. The tradeoff of number of grid points versus complexity of solution method is addressed. Requirements for high grid density in localized regions are also considered. The chapter concludes with a discussion of boundary condition sensitivities and the requirements for accurate discretization of impurity profiles including lateral extensions of source and drain regions in the simulation window.

Finite difference discretization on an irregular triangular grid is the subject of Chapter 3. The discretization and linearization of Poisson's equation with carrier statistics constraints is derived including the special case of discretization when the grid contains obtuse triangles. Discretization of the electron continuity equation is also described. The non-existence of an ex-



(a)



(b)

Fig. 1.5. Examples of PISCES graphical output (unretouched) showing equipotential lines for (a) MOSFET and (b) MESFET.

act two-dimensional equivalent to the one-dimensional Gummel-Scharfetter scheme is proved and a quasi-two-dimensional discretization is described.

Chapter 4 addresses methods for solving the discretized equations. Techniques such as the Fast Fourier Transform, the conjugate gradient method, and relaxation methods are discussed in connection with solving the matrix equations resulting from the discretization of Poisson's equation or the continuity equation. Methods for reducing the amount of required computation by renumbering the grid are examined. The various ways of solving the set of coupled equations are described with emphasis on the alternating method and its convergence properties. The convergence rate is shown to vary with device operating conditions and with carrier mobility. Several methods of accelerating the convergence of the alternating method are presented.

Chapter 5 presents two program application examples. The first is a typical application in device design in which punchthrough current is shown to vary greatly with a change in source/drain junction depth. The second application concerns mobility phenomena observed in strong inversion. Both of these applications demonstrate two important benefits of numerical analysis programs for semiconductor device design. First, these programs can be an aid in developing and/or proving theories about device behavior. Second, no *a priori* knowledge of device operating conditions are required. This is in contrast to the analytical modeling case where the proper analytical model must be chosen depending on the device region of operation (*e.g.* subthreshold, linear, breakdown).

The conclusions of this research and recommendations for further work are contained in Chapter 6.

## Chapter 2

### GRID

The execution time and storage requirements of two-dimensional semiconductor device simulation programs are directly dependent on the number of grid points (nodes) in the discretized analysis space. The number of equations to be solved is generally linearly related to the number of nodes and the number of arithmetic operations required for the solution is proportional to  $N^\alpha$  where  $N$  is the number of nodes and  $\alpha$  is somewhere between 1.5 and 2. Reducing the number of nodes in a simulation is, therefore, a matter of great importance.

#### 2.1 Grid Types

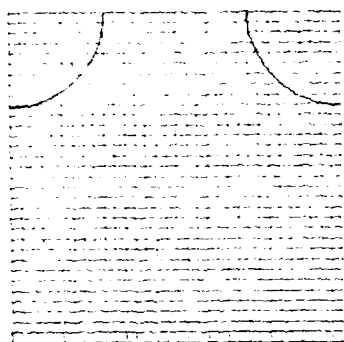
There are two types of grid which are of interest in two-dimensional device simulation: rectangular and triangular. Within each type there are variations which have substantial impact on the number of nodes and on the solution methods which can be used. Figure 2.1 shows several of the principal variations. Figure 2.1a shows a regular rectangular grid in which the grid spacing is constant, although not necessarily the same, in both the vertical and horizontal directions. This grid has the desirable feature that the discretization coefficients are constant in both directions so storage is minimized. It is also the grid on which nearly all numerical analysis theory is based, thus it was the grid used for most of the early work on device simulation. This grid is suitable for any of the solution methods discussed in the next chapter. Unfortunately, the semiconductor device problem requires

that the grid be very fine in some regions of the device, but not necessarily so in others. Therefore, the constant spacing rectangular grid wastes a lot of nodes in regions of the device where fine spacing is not required.

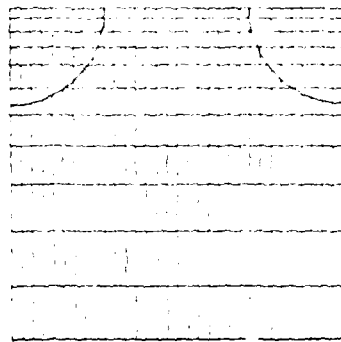
The semi-constant spacing rectangular grid of Figure 2.1b partially resolves this problem by allowing the grid to have variable spacing in one dimension. The savings in grid are not substantial, however, so this grid is of no great consequence except in its applicability to Fast Fourier Transform (FFT) solution techniques. The uniform grid spacing in the horizontal direction supports a fast solution to Poisson's equation through the use of the FFT. This technique, its advantages and limitations, will be discussed further in Chapter 4.

The most common grid is the variable spacing rectangular grid of Figure 2.1c. The grid spacing is allowed to vary in both directions, yet the coefficient storage required for an  $m$  by  $n$  grid is only on the order of  $m + n$ . The uniformity of the overall structure allows for simple, straight-forward, easy-to-program algorithms for equation solution regardless of the solution method used. The sole disadvantage of this grid is that it is still rather inefficient in grid allocation. Fine grid spacing at any point within the device results in grid lines which extend this spacing throughout the device in horizontal or vertical bands.

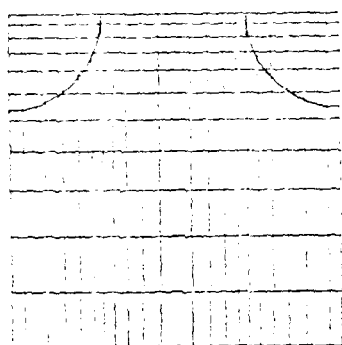
The grid of Figure 2.1d is a special case of a class of grids in which the grid lines are terminated within the simulation region of the device. The uniform horizontal spacing of this grid and the fact that the spacing remains uniform and exactly doubles as the vertical grid lines terminate means that this grid is also applicable to FFT solution techniques. A more generalized form of this grid has variable spacing in both directions and grid lines which may terminate in either direction. This is the type of grid used by Adler [2.1]



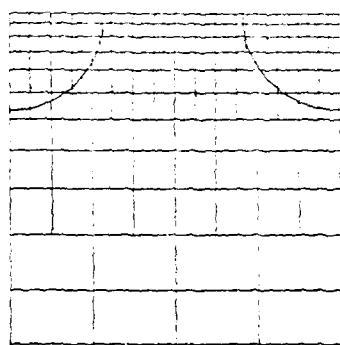
(a)



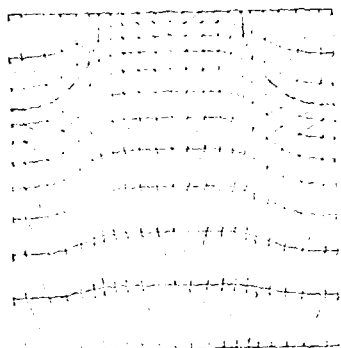
(b)



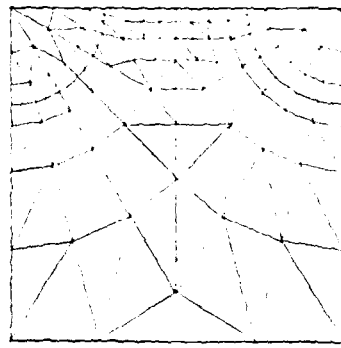
(c)



(d)



(e)



(f)

Fig. 2.1. Various rectangular and triangular grids.



in his simulation of thyristors. This grid is obviously flexible in its ability to place regions of coarse and fine grid throughout the device. The principal shortcoming of this type of grid is that as the number of terminating grid lines increases, most of the advantage of rectangular grids (i.e. small storage and simple algorithms) is lost and one may as well use a triangular grid.

The triangular grid of Figure 2.1e is based on a rectangular grid which has been distorted so that it conforms with features of the device - diagonals are added to divide each rectangle into two triangles. This grid has several desirable features. First, the underlying rectangular grid is easy for a user to specify. Moreover, it is not difficult to specify operations which distort the grid to the desired shape. Second, this grid retains its rectangular connectivity and thus supports simple solution methods such as line iterative techniques and maintains a well-defined matrix structure. On the negative side, each node has a unique set of discretization coefficients so that storage for an  $m$  by  $n$  grid is on the order of  $mn$ , the number of nodes. Another difficulty with this grid is that in regions of great distortion, the triangles may become unavoidably obtuse. This condition should be avoided if possible.

Finally, the completely general triangular grid of Figure 2.1f is the most efficient grid allocation in number of nodes. Spacing may be made arbitrarily fine or coarse in any local region with no global impact except that the transitions between regions should be gradual. Coefficient storage is again on the order of the number of nodes  $mn$ . The regular structure of all previous grids is lost, however, so that the solution techniques are necessarily less structured. The PISCES program was written assuming this type of grid; however, the grid generation portion of the program generates the more structured grid of Figure 2.1e.

The discussion of grid types to this point has assumed simulation of

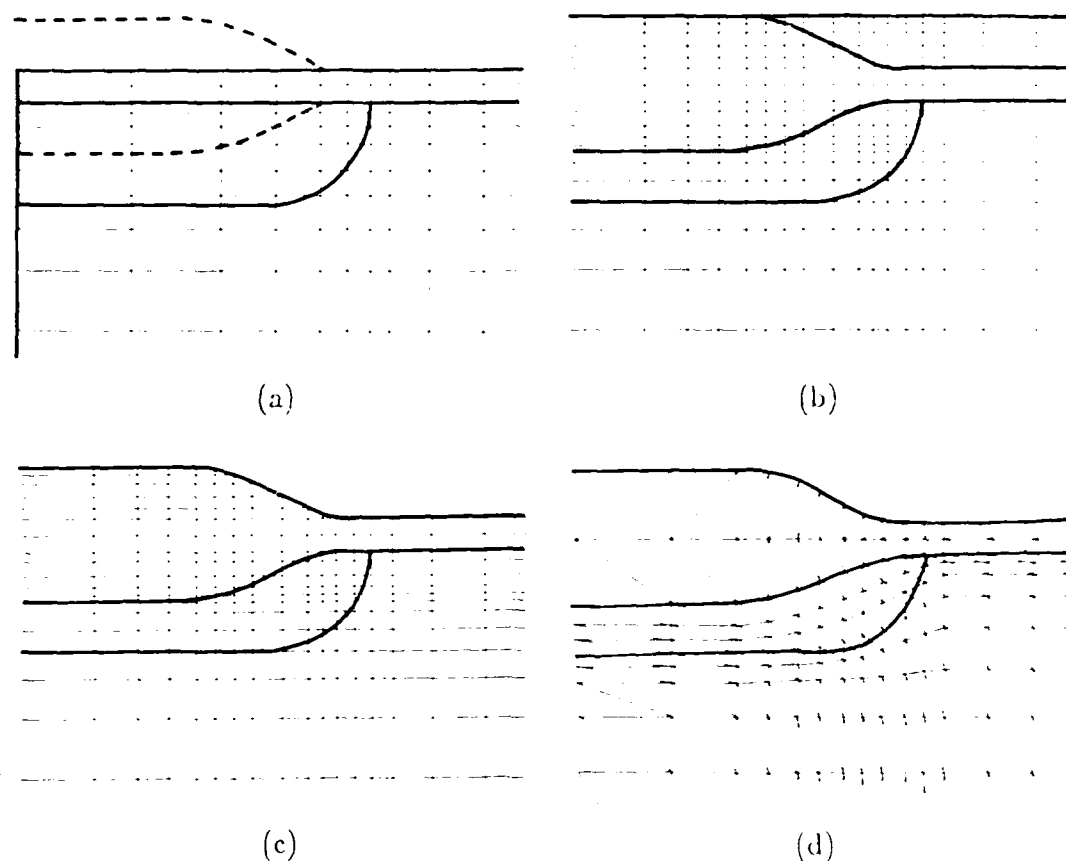


Fig. 2.2. Overlaying grid on non-rectangular structures.

a rectangular structure with regions of varying internal grid density. A more realistic case for semiconductor device simulation is to allow the device structure to be non-rectangular. This is mandatory for the modern two-dimensional devices described in Chapter 1. Figure 2.2 shows four methods of overlaying a grid on a non-rectangular structure. The structure shown is the source region of an IGFET with its overlying tapered oxide.

In Figure 2.2a the taper in the oxide is simply ignored and the device is assumed rectangular. The implication is that the effect of the portion of the device outside of the simulation region is insignificant. This generally

is not a good assumption although nearly all simulation programs using the finite difference method have used this type of simulation region. Making this assumption also requires that the insulator be modeled as a planar slab. Programs such as CADDET have taken advantage of this with the further assumption that the electric field in the insulator is one-dimensional. These assumptions simplify the simulation process and allow a more rapid, albeit less accurate, solution. Also, the FFT method mentioned earlier demands this type of grid overlay since it requires that the grid must be rectangular with uniform horizontal spacing. In addition, certain physical parameters such as permittivity must be constant along any horizontal grid line, limiting all material interfaces to be planar surfaces.

Figure 2.2b shows an overlay alternative in which the rectangular grid structure is maintained at the expense of wasting nodes which lie external to the device (i.e. in the space above the thin gate oxide). This appears to be a useful approach when using solution algorithms which depend on a rectangular simulation region. No application of this type of grid has been found in the literature. When the solution algorithms do not require a rectangular solution region, the nodes external to the device may be ignored, and the grid of Figure 2.2c results. This is the type of grid used by the GEMINI simulation program. Note that both grids (b) and (c) require an increased grid density at curved boundaries in order to accurately represent the boundary shape. Finally, the fully conforming triangular grid of Figure 2.2d provides the most accurate simulation with the minimum number of nodes.

## 2.2 Grid Density Criteria

Given a flexible discretization grid, one must then have a criteria for the proper placement of the grid points within the simulation region. In semiconductor device simulation, there are two principal driving factors: accurate representation of the potential and accurate representation of the net charge.

In the discretization of Poisson's equation, the assumption is made that the potential varies linearly between nodes (*i.e.* the electric field is constant), thus the grid spacing must be made sufficiently small in each direction so that a piecewise linear approximation to the true continuous potential is sufficiently accurate. This implies that the grid must be the finest in regions of high curvature of the potential. From Poisson's equation, it is easily seen that regions of high curvature of the potential correspond to regions of high net charge density,

$$\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}\right)\phi = \frac{-\rho}{\epsilon}.$$

For IGFETs, the net charge may be large in the surface inversion layer where there are large numbers of free carriers or in depletion regions where ionized impurities dominate. Vertical grid spacing in the inversion layer is typically  $.01 \mu\text{m}$  or less at the surface. Neutral regions, no matter how highly doped with impurities, do not generally require dense grid concentrations unless there is some chance that charge depletion or accumulation may occur there. The adequacy of a particular grid spacing for a given problem may be determined qualitatively by plotting the potential versus distance in the device and observing whether or not the piecewise nature of the potential approximation is evident. The same check may be performed quantitatively by comparing first and second order polynomial curve fits to the discrete

potential values.

It is also desirable that the total potential change between nodes not be too large. The criteria for potential changes between nodes depends on the device characteristics being simulated. For example, Figure 2.3 shows potential plotted versus lateral distance along the insulator-semiconductor surface for a device biased in punchthrough. The source is at the left, the channel region in the middle, and the drain at the right. For this bias condition, there is no inversion layer and conduction is impeded by the potential barrier near mid-channel. The barrier height is strongly controlled by the gate; however, it is also under control of the drain by virtue of the large drain bias. The drain current is proportional to  $e^{V_B/V_T}$  where  $V_B$  is the barrier height and  $V_T$  is the thermal voltage (approximately 26 millivolts). Thus, a 10 millivolt error in the simulated barrier height can result in an error in the simulated punchthrough current of nearly 50% depending on grid spacing and device profiles. As can be seen in the figure, the potential drops approximately 3 volts in .3  $\mu\text{m}$  near the drain; therefore, the horizontal grid spacing in this region must be fine enough that less than 10 millivolts error is made in discretizing the 3 volt drop. An error of 50% in subthreshold and punchthrough currents is typical for simulation programs due to this sensitivity. A similar situation exists for avalanche breakdown simulations where large voltage drops exist across short distances. In this case, however, one is generally most interested in the voltage at which breakdown occurs, not in accurate simulation of the current near breakdown. The voltage accuracy required is also not critical, being on the order of a few percent. The voltage drop per grid space may, therefore, be fairly large. Fortunately, currents are not as sensitive to voltage errors in the linear and saturation regions of operation where accurate current estimates are normally required.

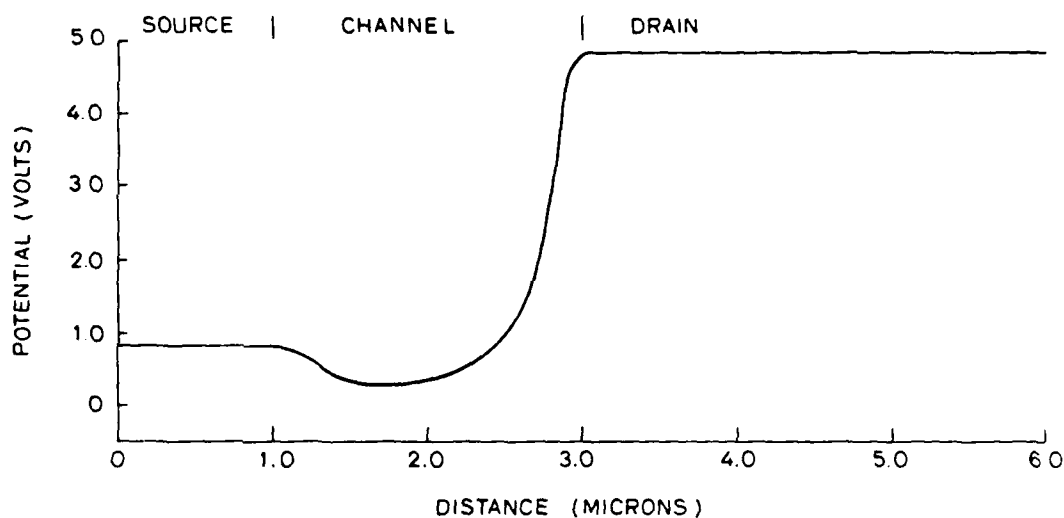


Fig. 2.3. Potential along the insulator-semiconductor interface for a device in punchthrough ( $V_G = -1V$ ,  $V_{BG} = 3V$ ,  $V_S = 0V$ ,  $V_D = 4V$ ).

If the region of high charge density is very thin then the change in potential across that region will be small in spite of the large curvature since potential is proportional to the second integral of charge. In this case, the requirement for accurate representation of net charge dictates the grid density required. This is typical of surface inversion layers where mobile charge densities may vary by several orders of magnitude over a distance of  $.1 \mu m$  or less. Since IGFET drain current is proportional to the integral of channel charge vertically from the insulator to the neutral bulk, it follows that one needs a finer grid in regions of high net charge concentration than in regions of low net charge. That is, a 10% error in mobile charge at a  $10^{19} \text{ cm}^{-3}$  concentration near the surface is much more significant than a 10% error in mobile charge at a  $10^{16} \text{ cm}^{-3}$  concentration away from the surface. The former would cause a nearly 10% error in drain current while

the latter would be insignificant. Actually, the difference is not this great. The effect is tempered somewhat by the fact that as the charge concentration decreases away from the surface, the gradient decreases also, spreading out the lower concentration regions. Thus the current carried by the thick low concentration regions away from the surface is more nearly equal to the current carried by the thin high concentration region near the surface.

In addition to inversion layers, accurate representation of net charge is also important near metallurgical junctions. Since the junctions will generally be depleted, the net impurity concentration is the driving factor. Although the net concentration of impurities at the junction is not high, the concentration gradients near the junction generally are. A fine grid is required, therefore, in order to accurately represent these steep gradients and to locate the junction. Figure 2.4 shows a PISCES grid with fine spacing normal to the surface in the channel region and normal to the junctions around the source and drain regions.

An additional point with regard to the accurate representation of charge concerns the allocation of impurities in a volume to the node representing that volume. Typically, the impurity distribution input to a device simulation program is evaluated at each discretization node and that value is assigned to the node. The total charge within that node's volume is the product of the volume and the assigned impurity concentration. In regions of low concentration gradients, this method is satisfactory; however, in regions of high concentration gradients, significant errors in total integrated charge may occur. A better method, therefore, is to integrate the input impurity distribution over the volume of the node then divide by the volume to arrive at an average impurity concentration. In this way, the total integrated impurity charge will be accurately represented. This method may allow larger grid

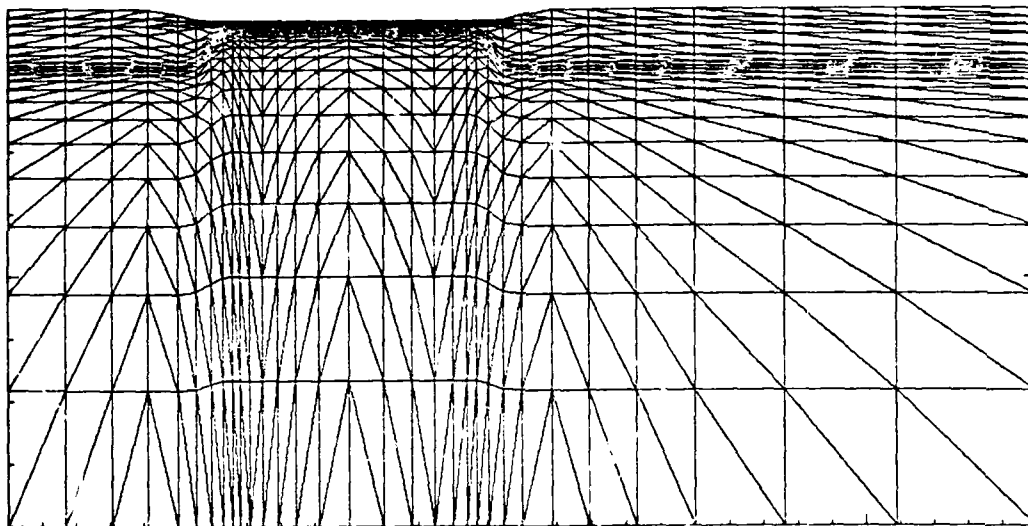


Fig. 2.4. PISCES grid for an IGFET.

spacing in the junction areas of the device.

### *2.3 Boundary Sensitivities*

One of the critical steps in device simulation is choosing the simulation region. This region must be chosen sufficiently large so that the active region of the device is accurately represented and is isolated from the deleterious effects of the simulation region boundary. In general, the simulation region must be large enough that any further increase in size has no effect on the results of the simulation. This is a check which may be used in practice. On the other hand, there is a competing desire to keep the simulation region small in order to reduce computation time and memory size. Also, the boundary conditions themselves must accurately represent the conditions at the



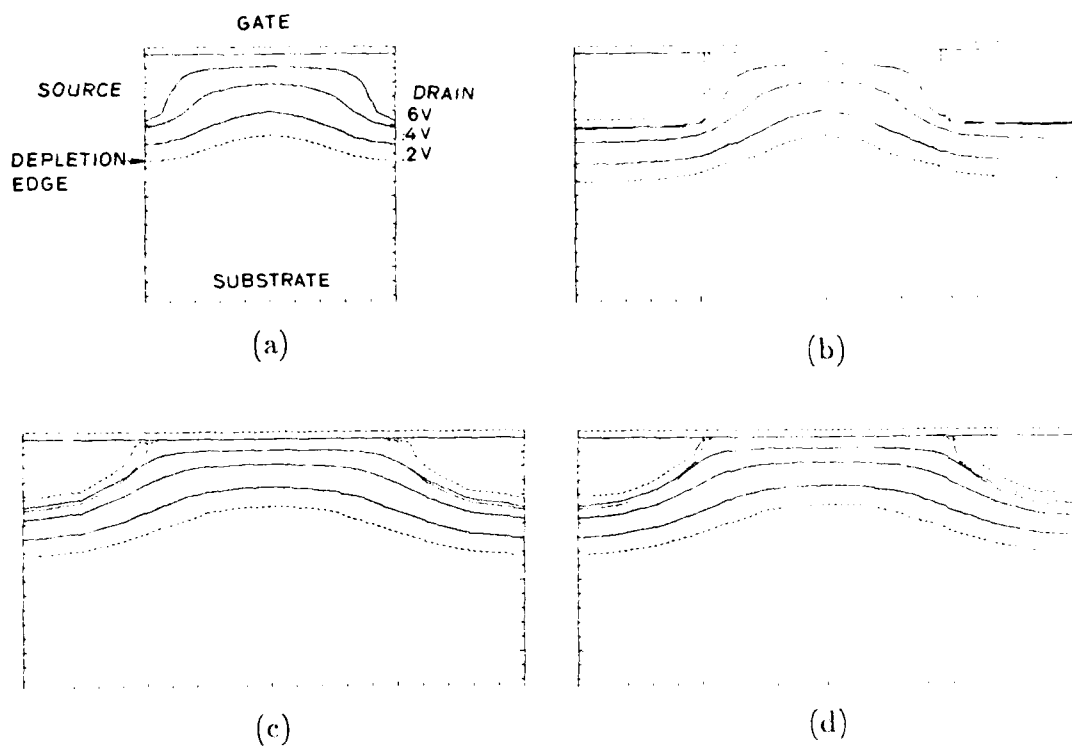


Fig. 2.5. Boundary condition sensitivities in the linear region;  $V_G = 2.4V$ ,  $V_{BG} = V_S = 0V$ , and  $V_D = .01V$ .

contacts to the device. Some device simulation programs make simplifying assumptions about these boundary conditions which greatly reduce the computation time at the expense of reduced accuracy in the solution.

Figure 2.5 shows the results of four PISCES simulations, each using a different boundary condition equivalent to those seen in other simulation programs. The simulation region was chosen to be rectangular in all four cases in order to maintain equivalency. The device simulated is an NMOS FET with a metallurgical channel length of  $2\mu m$ , gate length of  $4\mu m$  (except Figure 2.5a) substrate doping of  $10^{15} \text{ cm}^{-3}$  p-type and an oxide thickness of  $1000 \text{ \AA}$ .

In Figure 2.5a the source and drain regions are approximated by vertical

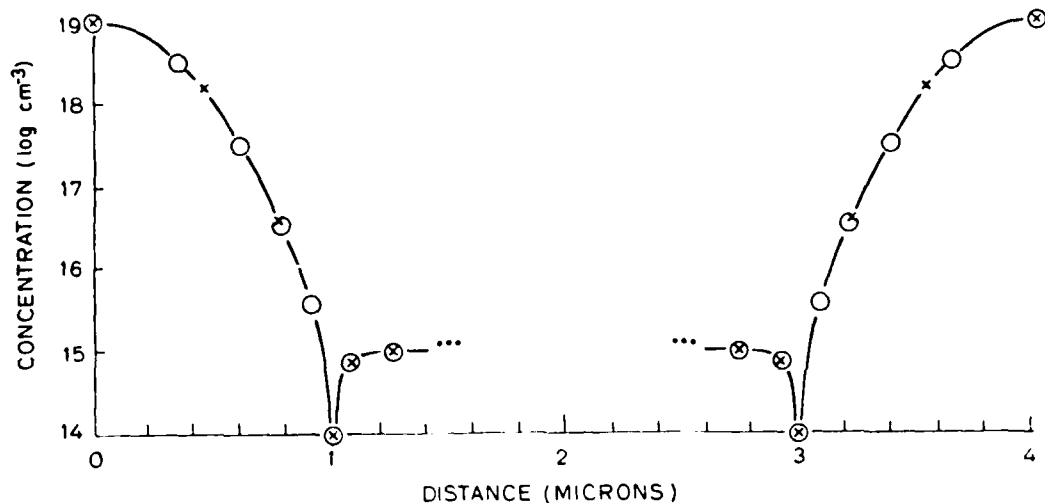


Fig. 2.6. Lateral impurity profile at the semiconductor surface.

boundary contacts  $1\text{ }\mu\text{m}$  deep as used in first generation simulation programs [2.2, 2.3]. The potential applied to this contact is reduced slightly from the external bias potential to account for the potential drop which occurs across the depletion region in the heavily doped side of an abrupt junction. In Figure 2.5b the source and drain regions are approximated by a uniformly doped rectangular region at  $10^{19}\text{ cm}^{-3}$ . Each region is  $1\text{ }\mu\text{m}$  deep and  $1\text{ }\mu\text{m}$  long. Figures 2.5c and 2.5d both have Gaussian source and drain profiles with peak concentrations of  $10^{19}\text{ cm}^{-3}$  and lateral and vertical junction depths of  $1\text{ }\mu\text{m}$ . The only difference in these two structures is in the accuracy of representation of the source/drain profiles. In the grid of Figure 2.5c, there are only four vertical grid lines in the source/drain while in the grid of Figure 2.5d there are six. The continuous doping profile and the grid spacings are shown in Figure 2.6.

In the direct contact case, Figure 2.5a, the equipotential lines are pulled

away from the surface at the edges. This means that the surface potential increases near the source and drain and the surface is more strongly inverted there than in mid-channel. Note also that the equipotential lines do not flatten out in the channel region but curve downwards along the entire channel length. This indicates that the long-channel assumption of a one-dimensional potential gradient in the channel does not hold and the device will exhibit short-channel characteristics (i.e. lowered threshold and increased drain conductance). The equipotential curves of the rectangular source/drain structure are nearly identical.

In contrast, the Gaussian source/drain structure of Figure 2.5c has flatter equipotential lines in the channel region which do not pull away from the surface as rapidly at the channel edges. Since the surface is less strongly inverted the drain current will be somewhat less. The equipotential lines of the finer source/drain grid spacing structure of Figure 2.5d are nearly identical.

The drain current for all four structures is plotted as a function of gate voltage in Figure 2.7. The currents were evaluated for gate voltages of .6V to 2.4V at .2V intervals which accounts for the piecewise continuous appearance of the plots. The characteristics of (a) and (b) are identical but differ from those of (c) and (d) which are themselves identical. The expected variation in drain current is seen to be approximately a factor of two at large gate biases but nearly an order of magnitude in the subthreshold region. The equivalence of the drain current for structures (c) and (d) implies that an accurate representation of the source/drain impurity profile may not be as important as representing its general shape, particularly in the vertical direction near the channel. Note, however, in Figure 2.6 that although the two grid spacings differ, both have a node exactly at the junction which serves to accurately

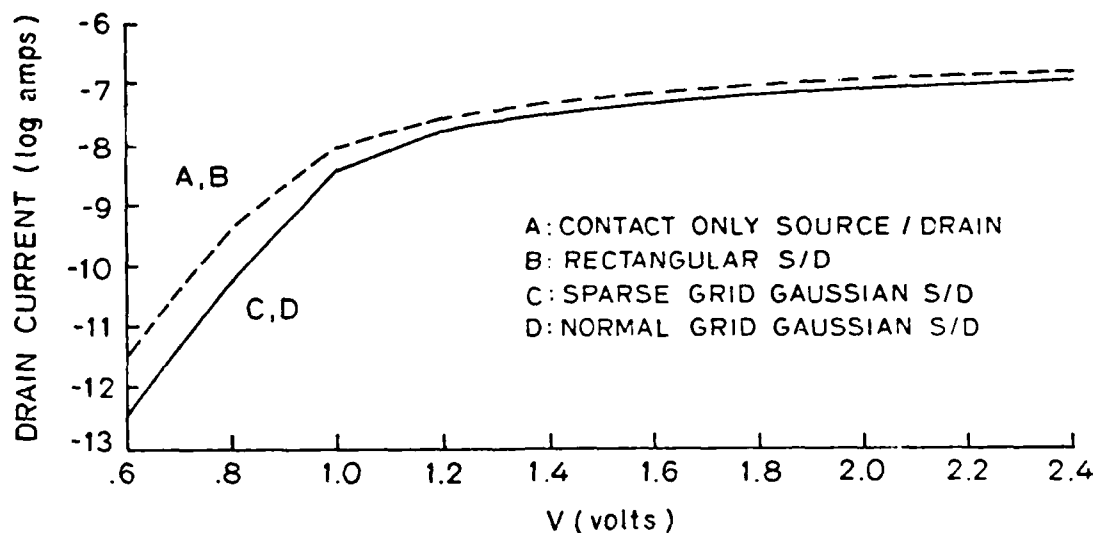


Fig. 2.7. Current sensitivity to boundary conditions in the subthreshold and linear regions for the structures of Figure 2.5.

locate the junction and thus fix the metallurgical channel length.

Figure 2.8 shows the same four structures biased in the saturation region of operation. The pinch-off point is clearly visible as the point along the channel where the equipotential lines become perpendicular to the insulator-semiconductor interface. To the left of this point, the electric field forces electrons toward the surface into the inversion layer. To the right of the pinch-off point, there is no inversion layer and the electric field tends to spread out the electrons as they travel toward the drain. In structures (a) and (b) the drain region pushes the equipotential lines down and to the left moving the pinch-off point to near mid-channel whereas in structures (c) and (d) it is much to the right of center. The distance from the source to the pinch-off point is the effective channel length and the channel length modulation caused by the pinch-off point moving to the left gives rise to drain conductance. Obviously,

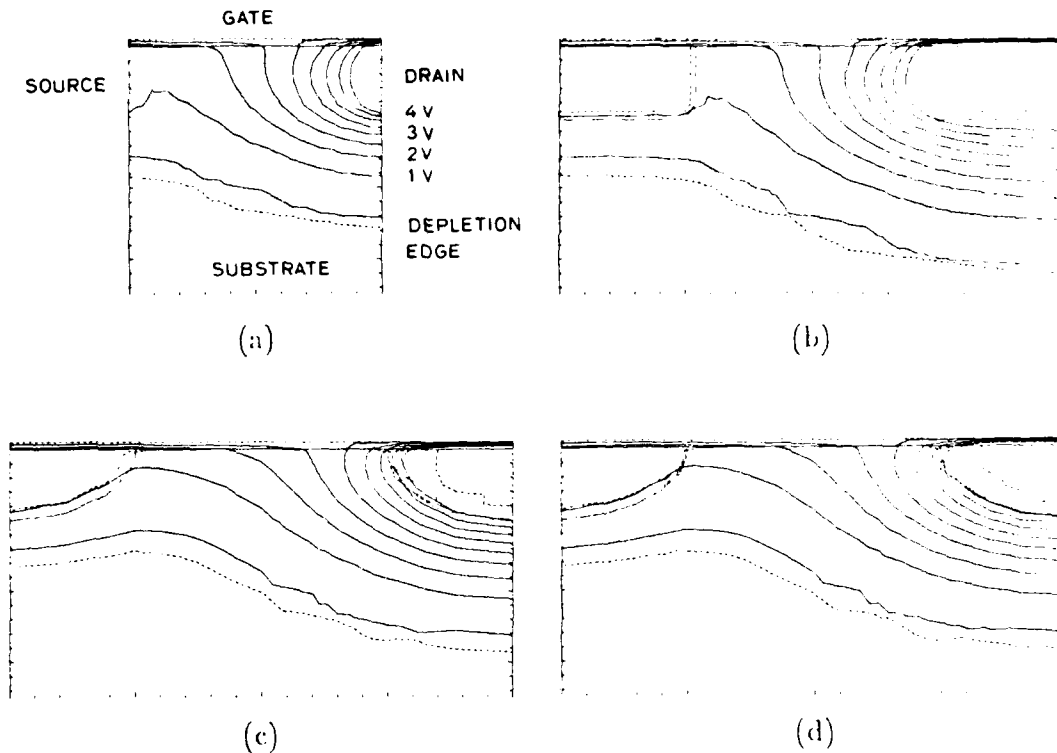


Fig. 2.8. Boundary condition sensitivities in the saturation region:  $V_G = 2V$ ,  $V_{BG} = V_S = 0V$ , and  $V_D = 5V$ .

in structures (a) and (b) the drain has greater control over the location of the pinch-off point; therefore, the currents are larger and the drain conductance is larger than for structures (c) and (d). This is illustrated in Figure 2.9 where drain current is plotted versus drain-to-source voltage at .5 volt increments. The gate voltage is 2 volts.

In addition to the larger currents and drain conductance expected for structures (a) and (b) it is also observed that there is a noticeable difference in current for (a) and (b) themselves. Close inspection of the equipotential line plots shows that indeed the pinch-off point for structure (b) is slightly to the left of that of structure (a) resulting in the larger current. The source of this slight shift, however, is obvious even in Fig 2.8. If the equipotential lines

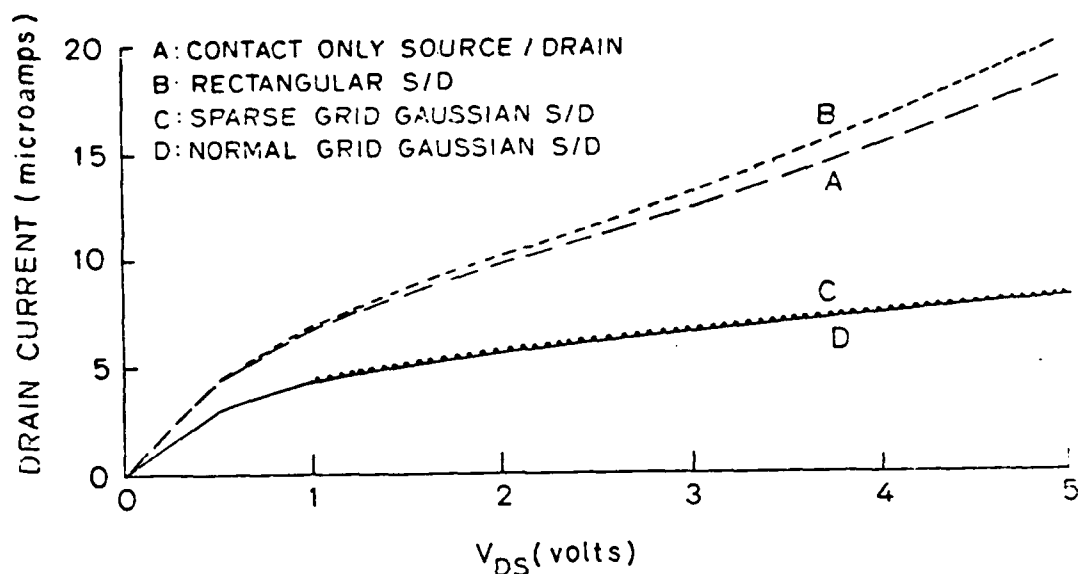


Fig. 2.9. Current sensitivity to boundary conditions in the saturation region for the structures of Figure 2.8.

directly underneath the left edge of the drain of structure (b) are compared to the equipotential lines underneath the drain contact of structure (a), the equipotential lines of structure (b) are seen to extend deeper into the substrate. This is because the left and right boundaries of the simulation region of structure (a) are too close to the active region of the device and the boundary condition assumptions are invalid. The contact portion of this boundary is a fixed potential boundary condition (Dirichlet) and is comparable to the rectangular source/drain assumption of structure (b). Along the remainder of the boundary of structure (a), however, a reflecting boundary condition (Neumann) assumption is made which is clearly inaccurate. That is, the larger simulation region of structure (b) shows that the equipotential lines are not at all symmetric about a vertical line below the left edge of the drain.

This illustrates the point made at the beginning of this section: the simulation region must be large enough to be independent of the solution. In particular, reflecting boundaries should only be placed where the problem is symmetric in some small region about the boundary. For IGFETs there are three boundaries to choose since the top surface is generally covered with the gate electrode which forms a natural boundary. The bottom boundary condition, whether Neumann or Dirichlet must be placed deeply enough into the device to be in a charge neutral region beyond any depletion regions which may arise in the simulation. Placement of the left and right boundaries is not as well defined. The Dirichlet portion of these boundaries representing the source and drain contact electrodes are not critical and may be placed anywhere within the neutral source/drain regions. The reflecting portions of these boundaries, however, dictate that they be placed sufficiently far from the channel region that the equipotential lines are naturally *one dimensional* (horizontal) in the vicinity of the boundary for all bias conditions simulated. Inspection of Figures 2.5 and 2.8 reveals that none of the simulations totally adhere to this rule especially along the drain boundary in the high drain bias case of Figure 2.8. The simulation region must contain some lateral extension of the one-dimensional portion of the source and drain regions and since the curvature of the equipotential lines extends out further at higher drain biases, the lateral extension of the drain region is significantly larger than that of the source side. Typically one to two channel lengths of extension on the drain side is sufficient for IGFETs at nominal drain biases. Note that since the potential and charge profiles in the lateral extension regions are almost one-dimensional, a coarse horizontal grid spacing is sufficient and little extra grid is required. An example of adequate lateral extension was shown in Figure 2.4.

## 2.4 Summary

The various types of analysis grids used for semiconductor device simulation are presented along with the advantages and disadvantages of each. Methods of overlaying these grids on non-rectangular structures are considered. The triangular grids are the most flexible in conforming to device shapes with the minimum number of nodes. It is shown that the density of the grid should vary in different regions of the device. The densest grid should occur in regions of high net charge density or large gradients of net charge. For IGFETs this means that the grid spacing should be small normal to inversion layers and metallurgical junctions. Checks for determining the adequacy of a particular grid spacing are suggested. The effects of large voltage drops between nodes are presented and shown to be highly problem dependent. An improved method for assigning impurity charge to a node is described which preserves the integrated net impurity charge.

Two of the source/drain boundary condition simplifications used by other device simulation programs have been examined and found to be grossly inaccurate. Sufficient grid to provide accurate location of the source/drain junctions along the channel is necessary but a coarse lateral spacing in the source/drain appears adequate as long as the vertical shape is accurately represented. The importance of proper choice of the simulation region has been demonstrated and suggestions are made for choosing this space.

The next chapter describes the finite difference discretization of the model equations on a triangular grid.



## Chapter 3

### DISCRETIZATION

Discretization of the semiconductor model equations using finite-difference techniques on rectangular grids is a well established procedure. Recent work by Greenfield [3.1] has expanded these procedures to show how to perform finite-difference discretization of Poisson's equation on a rectangular grid with non-planar surfaces and interfaces. The use of finite-difference techniques for discretization of the semiconductor model equations on a triangular grid, however, has not been previously described. The semiconductor model equations which describe the behavior of semiconductor devices are presented in what follows.

Poisson's equation describes the behavior of electric flux density in regions of net charge. Since charges are sources of electric flux, the flux density must diverge in regions of net charge as given by

$$\vec{\nabla} \cdot \vec{D} = \rho \quad (3.1)$$

where  $\vec{D}$  is the electric flux density and  $\rho$  is the net charge concentration.

The current continuity equation describes the time rate of change of carrier concentration. This concentration must change if there is not a balance between carrier generation, recombination, flux into, and flux out of a region of the device. This balance is expressed as

$$\frac{\partial n}{\partial t} = G_n - R_n + \frac{1}{q} \vec{\nabla} \cdot \vec{J}_n \quad (3.2a)$$

$$\frac{\partial p}{\partial t} = G_p - R_p - \frac{1}{q} \vec{\nabla} \cdot \vec{J}_p \quad (3.2b)$$

where  $n$  is the free electron concentration,  $G_n$  and  $R_n$  are the electron generation and recombination rates,  $q$  is the unit charge, and  $\vec{J}_n$  is the electron current density. The free hole concentration is  $p$  and the other quantities related to holes are analogous to those for electrons.

Carrier transport is described by a drift term dependent upon the electric field and a diffusion term dependent on the carrier concentration gradient. This is given by

$$\vec{J}_n = q\mu_n n \vec{E} + qD_n \vec{\nabla} n \quad (3.3a)$$

$$\vec{J}_p = q\mu_p p \vec{E} - qD_p \vec{\nabla} p \quad (3.3b)$$

where  $\mu_n$  is the electron mobility,  $\vec{E}$  is the electric field,  $D_n$  is the electron diffusion constant and  $\mu_p$  and  $D_p$  are the hole mobility and diffusion constant.

In nondegenerate semiconductors, the free carriers have a Boltzmann distribution of energy which leads to a relationship between the carrier concentration and potential as given by

$$n = n_I e^{q(\psi - \phi_n)/kT} \quad (3.4a)$$

$$p = n_I e^{q(\phi_p - \psi)/kT} \quad (3.4b)$$

where  $n_I$  is the intrinsic carrier concentration,  $\psi$  is the electrostatic potential,  $\phi_n$  and  $\phi_p$  are the electron and hole quasi-Fermi potentials,  $k$  is Boltzmann's constant and  $T$  is the absolute temperature. The nondegenerate assumption is not required for this work, but it allows simplification and provides a clearer picture of the discretization. Allowing degeneracy would require use of the Fermi-Dirac distribution function which leads to a Fermi-Dirac integral for the carrier concentration instead of the exponential. Numerical approximation of the Fermi-Dirac integral is computationally no more expensive than evaluation of the exponential, so little penalty is paid by allowing degeneracy.

Equations (3.1)-(3.4) represent the semiconductor model. The following relations serve to tie these equations together. The net charge  $\rho$  has three components

$$\rho = q(p - n + N) \quad (3.5)$$

where  $N$  is the net ionized impurity concentration. The electrostatic potential, electric field, and electric flux density are related by

$$\vec{E} = -\vec{\nabla}\psi \quad (3.6)$$

and

$$\vec{D} = \epsilon \vec{E} \quad (3.7)$$

where  $\epsilon$  is the permittivity. A link between the carrier mobility and diffusion constant is provided by the Einstein relation,

$$D_n = \frac{kT}{q} \mu_n \quad (3.8a)$$

$$D_p = \frac{kT}{q} \mu_p \quad (3.8b)$$

where again nondegeneracy is assumed. Allowing degeneracy here would require insertion of a multiplicative factor which is a function of the quasi-Fermi level. Finally, the thermal voltage is defined as

$$V_T \doteq \frac{kT}{q} \quad (3.9)$$

and the carrier generation and recombination terms are combined into a net recombination term as

$$U_n \doteq R_n - G_n \quad (3.10a)$$

$$U_p \doteq R_p - G_p \quad (3.10b)$$

in order to simplify the writing of equations.

Inserting the relations of Eqs. (3.5)-(3.10) into Equations (3.1) (3.4) and assuming steady-state such that  $\frac{\partial n}{\partial t} = \frac{\partial p}{\partial t} = 0$ , the semiconductor model equations become:

$$\vec{\nabla} \cdot (\epsilon \vec{\nabla} \psi) = -q(p - n + N) \quad (3.11)$$

$$\vec{\nabla} \cdot \vec{J}_n = qU_n \quad (3.12a)$$

$$\vec{\nabla} \cdot \vec{J}_p = -qU_p \quad (3.12b)$$

$$\vec{J}_n = q\mu_n(n\vec{E} + V_T\vec{\nabla}n) \quad (3.13a)$$

$$\vec{J}_p = q\mu_p(p\vec{E} - V_T\vec{\nabla}p) \quad (3.13b)$$

$$n = n_I e^{(\psi - \phi_n)/V_T} \quad (3.14a)$$

$$p = n_I e^{(\phi_p - \psi)/V_T} \quad (3.14b)$$

### 3.1 Poisson's Equation

Figure 3.1 shows a section of a hypothetical grid with five triangular sections labeled  $t_1$  -  $t_5$  having one common node variously labeled  $i_1$  -  $i_5$  and referred to as node  $i$ . The process of discretization involves the determination of two sets of parameters: the area assigned to each node, and the coupling coefficients between pairs of nodes.

The area assigned to a node is taken to be the area closer to that node than to any other node with which it shares a triangle. Thus, in the five triangle example of Figure 3.1, the area  $A_i$  bounded by the dashed line  $L_i$  represents the boundary of the area assigned to node  $i$ . This boundary is

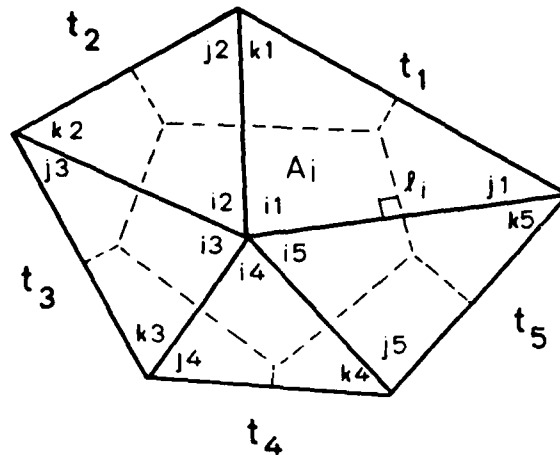


Fig. 3.1. Sample grid with five triangles. The triangles are labeled  $t_1$  -  $t_5$ ,  $A_i$  is the area associated with the central node, and  $l_i$  is the boundary of that area.

conveniently formed by the perpendicular bisector of each edge common to node  $i$ .

The coupling coefficients are obtained from the discretization of Poisson's equation. This is achieved by applying Gauss' law to Eq. (3.1) and converting integrations into summations. Applying Gauss' law to Eq. (3.1),

$$\oint_{l_i} \vec{D} \cdot d\vec{l} = \int_{A_i} \rho dA. \quad (3.15)$$

Inserting the relation of Eq. (3.7) and recognizing that the integrals may be evaluated by parts common to each triangle,

$$\begin{aligned} \oint_{l_{i1}} \epsilon \vec{E} \cdot d\vec{l} + \oint_{l_{i2}} \epsilon \vec{E} \cdot d\vec{l} + \dots + \oint_{l_{i5}} \epsilon \vec{E} \cdot d\vec{l} = \\ \int_{A_{i1}} \rho dA + \int_{A_{i2}} \rho dA + \dots + \int_{A_{i5}} \rho dA \end{aligned} \quad (3.16)$$

where  $l_{im}$  and  $A_{im}$  are those portions of the node  $i$  boundary and area lying within triangle  $t_m$ .

At this point, two assumptions are made: the permittivity and the electric field are constant within a triangle. The first of these requires only that material boundaries lie along triangle edges, a requirement easily met with the flexible triangular grids. The assumption of constant electric field within a triangle is a natural result of the triangular discretization since the only unique interpolation function of three points in two dimensions is a linear function, and linear potential implies a constant electric field.

These assumptions allow the line integrals to be replaced by dot products. Furthermore, since the net charge assigned to a node is considered to be evenly distributed throughout its area,  $\rho$  is a spacial constant and may be taken out of the integral. These changes to Eq. (3.16) yield

$$\begin{aligned} \epsilon_1 \vec{E} \cdot (\vec{l}_{i1j1} + \vec{l}_{i1k1}) + \epsilon_2 \vec{E} \cdot (\vec{l}_{i2j2} + \vec{l}_{i2k2}) + \cdots + \epsilon_5 \vec{E} \cdot (\vec{l}_{i5j5} + \vec{l}_{i5k5}) \\ = \rho_i A_i \end{aligned} \quad (3.17)$$

where  $\vec{l}_{imjm}$  is the vector normal to the portion of the boundary  $l_i$  in triangle  $t_m$  which is perpendicular to the  $ij$  side. The magnitude of  $\vec{l}_{imjm}$  is equal to the length of the boundary segment and the positive direction is away from node  $i$ . The other boundary normal vectors are similarly defined.

Figure 3.2 shows the labeling convention for a sample triangle. The vectors  $\vec{u}_j$  and  $\vec{u}_k$  are the unit vectors in the  $ij$  and  $ik$  directions respectively. The subscript denoting the triangle number has been dropped in the figure and the boundary segments  $l_{ij}$  and  $l_{ik}$  have been relabeled with their length  $h_j$  and  $h_k$ . Equation (3.17) may then be written in summation notation as

$$\sum_{1 \leq m \leq M} \epsilon_m \vec{E}_m \cdot (h_{jm} \vec{u}_{jm} + h_{km} \vec{u}_{km}) = \rho_i A_i \quad (3.18)$$

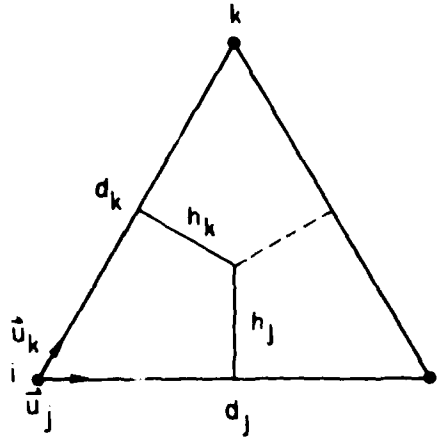


Fig. 3.2. Labeling convention for a triangle. The distance between nodes is  $d$ , the length (height) of the boundary segments is  $h$ , and  $\vec{u}$  is a unit vector.

where the problem has been generalized to include any number  $M$  of triangles containing the common node  $i$ .

The dot product  $\vec{E} \cdot \vec{u}_j$  is the component of the electric field in the  $ij$  direction. Due to the assumption of constant electric field within the triangle, this may be discretized using Equation (3.6) to obtain

$$\vec{E} \cdot \vec{u}_j = -\frac{\psi_j - \psi_i}{d_j} \quad (3.19)$$

where  $d_j$  is the distance between nodes  $i$  and  $j$ . Equation (3.18) may then be written as

$$\sum_{1 \leq m \leq M} \epsilon_m \left( (\psi_i - \psi_{jm}) \frac{h_{jm}}{d_{jm}} + (\psi_i - \psi_{km}) \frac{h_{km}}{d_{km}} \right) = \rho_i A_i. \quad (3.20)$$

This is the discretized form of Poisson's equation. Summing the terms over all triangles containing node  $i$  results in one equation containing the unknown

potentials of node  $i$  and all adjoining nodes. The term  $\frac{\epsilon_m h_{jm}}{d_{jm}}$  is referred to as the coupling coefficient between nodes  $i$  and  $j$  in triangle  $m$ . There is a similar coefficient for the same two nodes in the adjacent triangle (if there is one) which adds to this to form the full coupling coefficient. Note in Equation (3.20) that the coefficient of every node adjacent to node  $i$  is also a coefficient of node  $i$  with opposite sign, thus the coefficient of  $\psi_i$  is positive ( $\epsilon$ ,  $h$  and  $d$  are all positive) and exactly equal to the negative of the sum of the coefficients of all of the adjacent nodes.

Repeating the summation of Eq. (3.20) for each node in the grid results in a set of  $N$  equations in  $N$  unknowns where  $N$  is the number of nodes in the grid. The coefficient matrix for this set of equations is diagonally dominant. Note that if the grid is rectangular (properly composed of right triangles) the set of equations reduces to exactly that of the standard five-point difference scheme.

Unfortunately, the charge concentration  $\rho_i$  is a function of potential as described in Eq. (3.14). Combining Eqs. (3.5) and (3.14) for node  $i$  gives

$$\rho_i = q \left( n_I e^{(\phi_{pi} - \psi_i)/V_T} - n_{II} e^{(\psi_i - \phi_{ni})/V_T} + N_i \right), \quad (3.21)$$

thus the potential  $\psi_i$  appears non-linearly in the right hand side of Eq. (3.20). The resulting non-linear equation is solved using Newton's method. That is, it is linearized and solved iteratively until converged.

Combining equations (3.20) and (3.21), linearization is achieved by first replacing every  $\psi$  by  $\psi + \Delta\psi$ ,



$$\begin{aligned}
& \sum_{1 \leq m \leq M} \epsilon_m \left( ((\psi_i + \Delta\psi_i) - (\psi_{jm} + \Delta\psi_{jm})) \frac{h_{jm}}{d_{jm}} \right. \\
& \quad \left. + ((\psi_i + \Delta\psi_i) - (\psi_{km} + \Delta\psi_{km})) \frac{h_{km}}{d_{km}} \right) \\
& = qA_i \left( n_i e^{(\phi_{pi} - (\psi_i + \Delta\psi_i))/V_T} - n_i e^{((\psi_i + \Delta\psi_i) - \phi_{ni})/V_T} + N_i \right) \\
& = qA_i \left( p_i e^{-\Delta\psi_i/V_T} - n_i e^{\Delta\psi_i/V_T} + N_i \right) \tag{3.22}
\end{aligned}$$

then taking first order Taylor-series approximations to the exponentials,

$$\begin{aligned}
& = qA_i (p_i (1 - \Delta\psi_i/V_T) - n_i (1 + \Delta\psi_i/V_T) + N_i) \\
& = \rho_i A_i - qA_i (p_i + n_i) \Delta\psi_i/V_T. \tag{3.23}
\end{aligned}$$

Placing all terms in  $\Delta\psi$  on the left hand side results in the iterative form,

$$\begin{aligned}
& \sum_{1 \leq m \leq M} \epsilon_m \left( (\Delta\psi_i - \Delta\psi_{jm}) \frac{h_{jm}}{d_{jm}} + (\Delta\psi_i - \Delta\psi_{km}) \frac{h_{km}}{d_{km}} \right) \\
& \quad + qA_i (p_i + n_i) \Delta\psi_i/V_T \\
& = - \sum_{1 \leq m \leq M} \epsilon_m \left( (\psi_i - \psi_{jm}) \frac{h_{jm}}{d_{jm}} + (\psi_i - \psi_{km}) \frac{h_{km}}{d_{km}} \right) + \rho_i A_i. \tag{3.24}
\end{aligned}$$

The right hand side of Eq. (3.24) is the residual of Poisson's equation and the left hand side contains the unknown potentials for the Newton step. After each solution for the  $\Delta\psi$ 's, the potentials are updated, the potential dependent terms are re-evaluated, and the iteration proceeds until convergence is achieved.

When Eq. (3.24) is assembled for every point in the grid and put in matrix form, the coefficient matrix has exactly the same terms in it as the coefficient matrix of Eq. (3.20) except that positive terms resulting from the linearization of the charge concentration have been added to the entries on the diagonal. This matrix is positive definite resulting in desirable convergence properties.

### 3.1.1 Area Allocation

In the foregoing discussion, the vectors  $h_{jm}\vec{u}_{jm}$  and  $h_{km}\vec{u}_{km}$  of Eq. (3.18) were kept separate in order that the electric field may be expressed in terms of its components in the  $\vec{u}_{jm}$  and  $\vec{u}_{km}$  directions. As shown in Eq. (3.20) these components are conveniently obtained from the potential differences between nodes. Additional insight may be gained, however, by performing the vector sum as illustrated in Figure 3.3. The electric flux crossing the node  $i$  area boundary  $l_i$  within this triangle is identically the flux crossing the line segment  $s$  joining the midpoints of the adjacent sides, thus the choice of the node  $i$  area boundary is somewhat arbitrary. In fact, the above discretization applies to any simple boundary which has these midpoints as its endpoints. Four random possibilities are shown in Figure 3.4.

One such alternative boundary is to use the line segments joining the triangle centroid to the adjacent side midpoints in the manner of Winslow [3.2]. With this boundary, each node is allocated one-third of the total triangle area. This allocation scheme was tried with the PISCES program and generally yielded satisfactory results except that some assemblies of triangles resulted in uneven distribution of area. An example is shown in Figure 3.5 in which two similar discretization grids are given with the only difference being the triangle orientation. Note that in case (a) there are four identical triangles connected to the central node while in case (b) there are eight. As a result, the central node is allocated twice as much area in case (b) as in case (a). This unequal weighting results in perturbations to the desired solution. In contrast, application of the perpendicular bisector method to the grids of Figure 3.5 results in equal allocation of area for the two cases. The perpendicular bisector method of forming the node area boundary appears

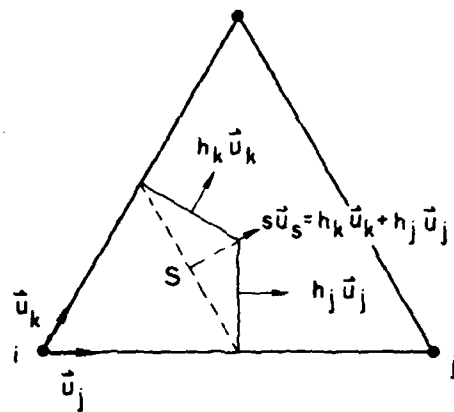


Fig. 3.3. Equivalence of flux boundary  $s$  to flux boundaries  $h_j$  plus  $h_k$ .

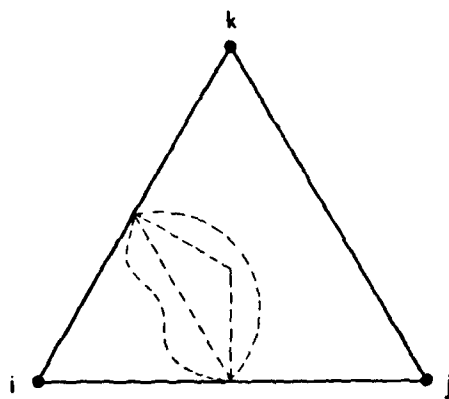


Fig. 3.4. Various boundaries with equivalent flux conservation characteristics.

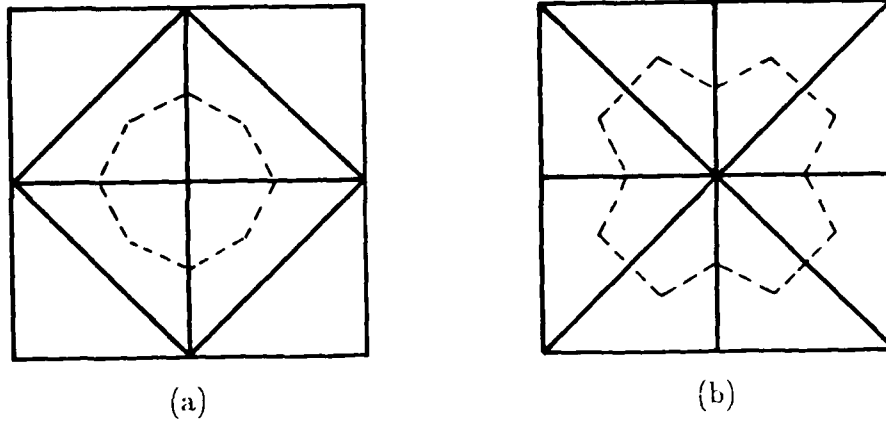


Fig. 3.5. Dependence of area weighting on triangle orientation using the centroid method. The area allocation in case (b) is twice that of case (a).

to be the best area allocation scheme for triangular grids in semiconductor problems.

### 3.1.2 Obtuse Triangles

The above derivations work quite nicely for acute triangles; however, when the triangles become obtuse, adjustments need to be made. The need arises from the fact that the intersection of the perpendicular bisectors of the sides of an obtuse triangle occurs outside of the triangle as illustrated in Figure 3.6. In this example, the coupling coefficients are exactly as derived earlier except that the coupling coefficient between nodes  $i$  and  $j$  becomes negative,  $-\frac{\epsilon h_j}{d_j}$  since the segment  $h_j$  has reversed direction. It can be shown that this choice of coupling coefficients still accurately accounts for the flux passing through the line segment  $s$ . Intuitively, it may be reasoned that the vector sum of the directed line segment  $h_k$  minus the directed line segment  $h_j$  is the directed line segment  $s$ , so the vector sum of their normals also equate.

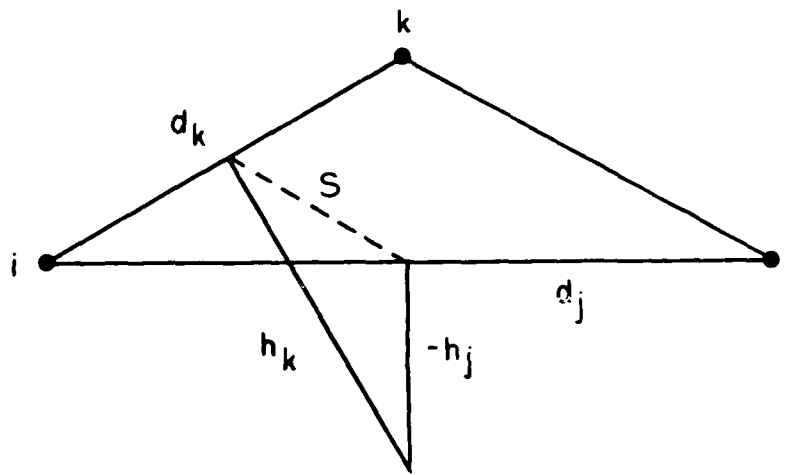


Fig. 3.6. Coupling coefficients for Poisson's equation on an obtuse triangle.

Note that the coupling coefficient goes from positive to zero to negative as the triangle goes from acute to right to obtuse, so there is no discontinuity in the transition from acute to obtuse.

The total coupling coefficient for nodes  $i$  and  $j$  includes a contribution from both triangles which share the side joining nodes  $i$  and  $j$ . Thus, although one component of the coupling coefficient may be negative, the total coupling coefficient may still be positive if the contribution from the adjacent triangle is sufficiently large and positive. It can be shown that the total coupling coefficient will be positive if the sum of the opposite angles is less than 180 degrees. This condition can always be satisfied for triangles which are not on the solution region boundary or on a material interface by reconnecting the triangles. That this is so can easily be proven by hypothesizing a situation in which two adjacent triangles exist whose angles opposite the common edge sum to more than 180 degrees. If the four nodes composing these two triangles are looked upon as a quadrilateral, then the common edge is a diagonal of

the quadrilateral, the sum of whose interior angles must equal 360 degrees. If instead of dividing the quadrilateral into two triangles using this diagonal, the other diagonal is used, two triangles result whose angles opposite their common side necessarily sum to less than 180 degrees. Obviously, if the edge in question composes a solution region boundary or a material interface, the triangles may not be reconstructed; however, subdivision of a triangle into two or more triangles can eliminate the problem in these special cases.

The occurrence of obtuse angles also complicates the area allocation method. Looking again at Figure 3.6, the boundary for the area allocated to node  $i$  is no longer defined by the line segments  $h_j$  and  $h_k$  as they were in Figure 3.2. One possible choice of boundary is that portion of line segment  $h_k$  which lies within the triangle. Unfortunately, this choice does not meet the prerequisite that the boundary include the midpoints of the adjacent sides, thus there would not be conservation of flux within the triangle using this boundary.

A better choice for the boundary of the area allocated to node  $i$  is the line segment  $s$  in Figure 3.6, the segment joining the midpoints of the adjacent sides. This choice satisfies the condition for conservation of flux and is somewhat better than the centroid method in area weighting in that it allocates one-fourth of the triangle area to node  $i$ , one-fourth to node  $j$ , and one-half to node  $k$ . The acute and obtuse area allocation schemes are identical for an angle of 90 degrees so there is a smooth transition from one to the other. Figure 3.7 shows a hypothetical grid with various triangle types to demonstrate the area allocation scheme. The solid lines are the triangular grid and the dashed lines are the area boundaries.

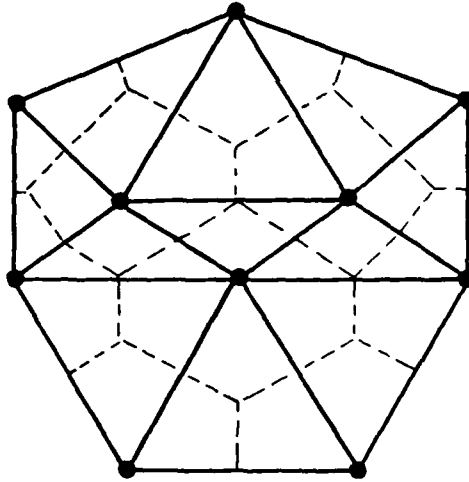


Fig. 3.7. Area allocation with a mix of acute and obtuse triangles.

### 3.2 Continuity Equation

The discretization of the continuity equation will be described only for electrons since the discretization for holes is analogous with the exception of signs. Also, the discretization of the electron transport equation will be performed in one dimension only. The reason for this will be explained later.

#### 3.2.1 Electron Transport Equation

From Eq. (3.13), the electron transport equation in one dimension is

$$J_n = q\mu_n \left( nE + V_T \frac{\partial n}{\partial x} \right). \quad (3.25)$$

The standard finite-difference approach to discretization of this equation would result in

$$J_n(\Delta x) = q\mu_n(\Delta x) \left( n(\Delta x)E(\Delta x) + V_T \frac{n(\Delta x) - n(0)}{\Delta x} \right) \quad (3.26)$$

where a uniform spacing of  $\Delta x$  is assumed for simplicity. Insertion of Eq. (3.28) into the continuity equation (3.12a) yields a single equation in the unknown  $n$  assuming that the electric field is given. However, it can be shown that this form of the transport equation leads to numerical instability whenever the voltage change between nodes exceeds  $2V_T$ . This point was noted by Scharfetter and Gummel [3.3] and has resulted in formulation of a discretization scheme which avoids this difficulty. The method will be summarized here in order to provide needed background for further discussion.

Equation (3.25) may be rearranged as

$$\frac{\partial n}{\partial x} + \frac{E}{V_T} n = \frac{J_n}{q\mu_n V_T}. \quad (3.27)$$

Given two nodes separated by a distance  $d$  as in Figure 3.8 and assuming that  $J_n$ ,  $\mu_n$ , and  $E$  are constant between these nodes, then Eq. (3.27) may be viewed as a first order differential equation in  $n$  with constant coefficients. The general solution to this equation is

$$n = Ce^{-Ex/V_T} + \frac{J_n}{q\mu_n E}. \quad (3.28)$$

Evaluating this equation for  $n = n_i$  at  $x = 0$  results in

$$C = n_i - \frac{J_n}{q\mu_n E}. \quad (3.29)$$

Using this value for  $C$  and evaluating Eq. (3.28) at node  $j$  yields

$$n_j = \left( n_i - \frac{J_n}{q\mu_n E} \right) e^{-Ed/V_T} + \frac{J_n}{q\mu_n E} \quad (3.30)$$

which may be rearranged as

$$J_n = q\mu_n E \left( \frac{n_j}{1 - e^{-Ed/V_T}} + \frac{n_i}{1 - e^{Ed/V_T}} \right) \quad (3.31)$$



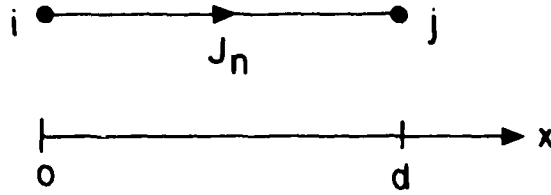


Fig. 3.8. Labeling convention for one-dimensional current transport.

or equivalently

$$J_n = q\mu_n \frac{(\psi_i - \psi_j)}{d} \left( \frac{n_j}{1 - e^{-(\psi_i - \psi_j)/V_T}} + \frac{n_i}{1 - e^{(\psi_i - \psi_j)/V_T}} \right). \quad (3.32)$$

This is the Scharfetter-Gummel form for discretization of the electron transport equation in one dimension. One may easily verify that in the limit as the potential difference between nodes approaches zero, Eq. 3.32 becomes identically the diffusion term while as the potential difference becomes large, it becomes identically the drift term.

### 3.2.2 Electron Continuity Equation

The discretization of the electron continuity equation will be performed in a quasi-two-dimensional form using the one-dimensional transport equation. As with Poisson's equation, we begin with the acute triangle case and use the boundaries shown in Figures 3.1 and 3.2. Applying Gauss' Theorem to Eq. (3.12a),

$$\oint_{\Lambda_i} \vec{J}_n \cdot d\vec{l} = \int_{\Lambda_i} qU_n d\Lambda. \quad (3.33)$$

The discretization process from this point parallels that of the Poisson equation (Eq. 3.15 to Eq. 3.20) except that quantities on the left hand side of the equation are taken to be constant only along an edge of the triangle instead of over the entire triangle. The discretization leads to a form which is similar to that of Eq. (3.18)

$$\sum_{1 \leq m \leq M} \left( \vec{J}_{njm} \cdot (h_{jm} \vec{u}_{jm}) + \vec{J}_{nkm} \cdot (h_{km} \vec{u}_{km}) \right) = qU_n A_i \quad (3.34)$$

where  $\vec{J}_{njm}$  is the electron current density along edge  $ij$  in triangle  $m$  and  $\vec{J}_{nkm}$  is similarly defined. Note that since  $\vec{J}_n$  cannot be assumed constant over the triangle,  $\vec{J}_{njm} \neq \vec{J}_{nkm}$ . Since  $\vec{J}_{njm}$  and  $\vec{J}_{nkm}$  are one-dimensional current density vectors in the  $\vec{u}_{jm}$  and  $\vec{u}_{km}$  directions respectively, the dot products become multiplies and Eq. 3.34 becomes

$$\sum_{1 \leq m \leq M} (J_{njm} h_{jm} + J_{nkm} h_{km}) = qU_n A_i \quad (3.35)$$

where the current density is taken to be positive in the direction away from node  $i$ . Thus the electron current out of the area of node  $i$  into the area of node  $j$  in triangle  $m$  is the current density,  $J_{njm}$ , times the length of the perpendicular boundary through which the current flows,  $h_{jm}$ . This same current density also flows through a perpendicular segment of the node  $i$  boundary in the adjacent triangle. The total electron current out of the node  $i$  area is the summation of the current crossing the area boundary into each adjacent node, and this total must equal the recombination rate  $qU_n A_i$ .

Finally, Eq. (3.32) may be inserted into Eq. (3.35) dropping the  $q$  from both sides to yield

$$\begin{aligned}
& \sum_{1 \leq m \leq M} \left( \frac{\mu_{njm} h_{jm}}{d_{jm}} (\psi_i - \psi_{jm}) \left( \frac{n_{jm}}{1 - e^{-(\psi_i - \psi_{jm})/V_T}} + \frac{n_i}{1 - e^{(\psi_i - \psi_{jm})/V_T}} \right) \right. \\
& \quad \left. + \frac{\mu_{nkm} h_{km}}{d_{km}} (\psi_i - \psi_{km}) \left( \frac{n_{km}}{1 - e^{-(\psi_i - \psi_{km})/V_T}} + \frac{n_i}{1 - e^{(\psi_i - \psi_{km})/V_T}} \right) \right) \\
& = U_n A_i.
\end{aligned} \tag{3.36}$$

This is the fully discretized electron continuity equation. The unknowns are the electron concentrations  $n_i$ ,  $n_{jm}$ , and  $n_{km}$  while the mobilities, potentials, and electron recombination rate are assumed to be known; although, each is actually a function of the electron concentration. The potentials are a strong function of the electron concentration through Poisson's equation, thus alternating solutions of the Poisson and continuity equations are required until convergence is obtained. The mobilities and electron recombination rates may also be functions of the electron concentrations, but the functional relationships are generally very weak so that merely updating the mobility and recombination rate after new electron concentrations are computed is sufficient.

When Eq. (3.36) is assembled for every node in the grid, one again has  $N$  equations in  $N$  unknowns as in the Poisson case; however, the continuity equation does not have to be solved in the insulator regions of the device so that one may limit the solution to only those nodes lying in the semiconductor region. The coefficient matrix does not have the desirable iteration properties of the Poisson coefficient matrix, but since iteration is not required to solve the continuity equation, this point is not critical. Also, the coefficients look rather formidable at first, but a careful look will show that the coefficients are well behaved and, in fact, no difficulties in obtaining a solution were ever observed in the PISCLS program. The well behaved nature of the coefficients is demonstrated in the fact that

$$\frac{\psi_1 - \psi_2}{1 + e^{(\psi_1 - \psi_2)/V_T}} \approx \begin{cases} 0, & \text{for } \frac{\psi_1 - \psi_2}{V_T} \gg 0; \\ -V_T, & \text{for } \frac{\psi_1 - \psi_2}{V_T} \approx 0; \\ \psi_1 - \psi_2, & \text{for } \frac{\psi_1 - \psi_2}{V_T} \ll 0. \end{cases} \quad (3.37)$$

Finally, note that the coupling terms  $\frac{\mu_n h}{d}$  in Eq. (3.36) are analogous to the terms  $\frac{\epsilon h}{d}$  in Poisson's equation, Eq. (3.20).

### 3.2.3 Obtuse Triangles

As in the case of Poisson's equation, adjustments need to be made to the continuity equation discretization in obtuse triangles. In the Poisson discretization, the node area boundary was pinned at the midpoints of the sides as the triangle became obtuse while the coupling coefficients,  $\frac{\epsilon h}{d}$ , were allowed to become negative in a rather natural manner. The continuity equation necessarily uses the same boundary as Poisson's equation but cannot allow its  $\frac{\mu_n h}{d}$  terms to become negative as the triangle becomes obtuse. This discrepancy stems from the quasi-two-dimensional nature of the continuity equation discretization.

Figure 3.9 graphically depicts the reasoning behind the choice of coupling coefficients for the continuity equation in acute and obtuse triangles. The current density flowing from node  $a$  node  $b$  is shown as  $\vec{J}_n$ . In triangle  $t_2$  this current density passes from the node  $i_2$  area into the node  $j_2$  area through a cross-sectional window of width  $h_{j2}$  which is the length of the boundary segment. (Note that node  $a$  is equivalent to both  $i_1$  in  $t_1$  and  $i_2$  in  $t_2$ ). However, in the obtuse triangle,  $t_1$ , the current density passes through a cross-sectional width of  $h_{k1}$  which is the projected length of the boundary segment onto the perpendicular to the current density vector. The value of

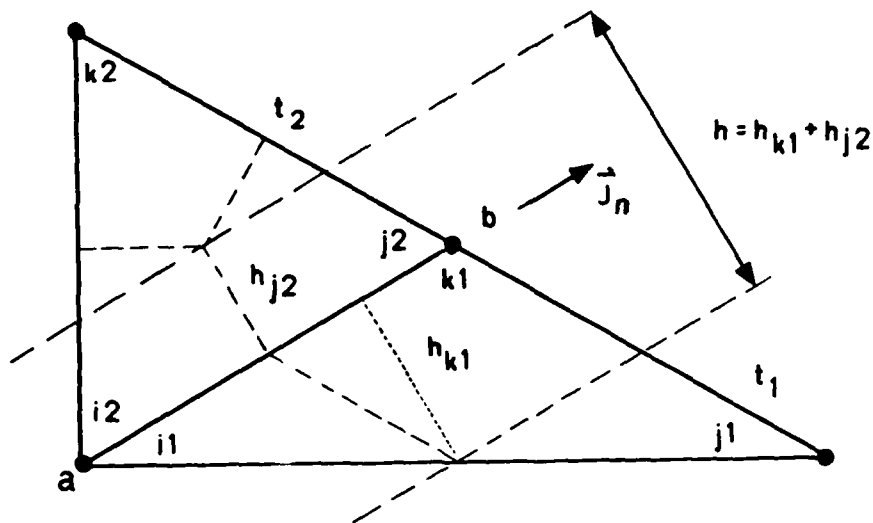


Fig. 3.9. Coupling coefficient derivation for the continuity equation on acute and obtuse triangles.

$h$  used in the total coupling coefficient between nodes  $a$  and  $b$  is the sum of these two terms, *i.e.* the total cross-sectional width of the common boundary between the two nodes.

Note in obtuse triangle  $t_1$ , that there is no common boundary between nodes  $i_1$  and  $j_1$  so  $h_{j1} = 0$  and no current is allowed to flow between them. Current may be allowed along this same edge in an adjacent triangle, however, if the opposite angle is acute. This condition can be assured, as was shown in Section 3.1.2, as long as the triangle edge is not a solution region boundary or a material interface. Even these cases can be handled by subdivision of the triangles, so it is never necessary for the continuity equation coupling coefficient between two adjacent semiconductor nodes to be zero (except in the degenerate case of a perfectly rectangular grid structure where the triangle edges representing the diagonals of the rectangles have coefficients of zero).

Note also that as in the Poisson discretization, there is no discontinuity in the choice of coupling coefficients as the triangle passes from acute to obtuse.

### 3.2.4 Absence of Two-Dimensional Scharfetter-Gummel Discretization

As mentioned earlier, the Scharfetter-Gummel algorithm for discretization of the electron transport equation does not have an analogous form in two dimensions. This can be shown as follows. Equation (3.13a) may be rewritten in two dimensional form as

$$\begin{aligned}\vec{J}_n &= q\mu_n \left( n(E_x \vec{u}_x + E_y \vec{u}_y) + V_T \left( \frac{\partial n}{\partial x} \vec{u}_x + \frac{\partial n}{\partial y} \vec{u}_y \right) \right) \\ &= q\mu_n \left( nE_x + V_T \frac{\partial n}{\partial x} \right) \vec{u}_x + q\mu_n \left( nE_y + V_T \frac{\partial n}{\partial y} \right) \vec{u}_y \\ &= J_{nx} \vec{u}_x + J_{ny} \vec{u}_y\end{aligned}\quad (3.38)$$

where  $\vec{u}_x$  and  $\vec{u}_y$  are the unit vectors in the  $x$  and  $y$  directions. Making the necessary assumptions that  $\mu_n$ ,  $\vec{J}_n$  and  $\vec{E}$  are constant over the triangle, then the  $x$  and  $y$  components of Eq. (3.38) may be handled separately using the one-dimensional algorithm. In the  $x$  direction,

$$\begin{aligned}\frac{\partial n}{\partial x} &= -\frac{E_x}{V_T} n + \frac{J_{nx}}{q\mu_n V_T} \\ &= \alpha n + \beta\end{aligned}\quad (3.39)$$

where  $\alpha$  and  $\beta$  take on the obvious values. Similarly, in the  $y$  direction,

$$\begin{aligned}\frac{\partial n}{\partial y} &= -\frac{E_y}{V_T} n + \frac{J_{ny}}{q\mu_n V_T} \\ &= \gamma n + \delta.\end{aligned}\quad (3.40)$$

It is a necessary and sufficient condition for a solution to exist for Eqs. (3.39) and (3.40) that

$$\frac{\partial^2 n}{\partial x \partial y} = \frac{\partial^2 n}{\partial y \partial x} \quad (3.41)$$

Evaluating these terms yields

$$\begin{aligned}\frac{\partial^2 n}{\partial x \partial y} &= \frac{\partial}{\partial x}(\gamma n + \delta) \\ &= \gamma(\alpha n + \beta)\end{aligned}\tag{3.42}$$

and

$$\begin{aligned}\frac{\partial^2 n}{\partial y \partial x} &= \frac{\partial}{\partial y}(\alpha n + \beta) \\ &= \alpha(\gamma n + \delta).\end{aligned}\tag{3.43}$$

Thus, the necessary and sufficient condition for a solution to exist is

$$\gamma \alpha n + \gamma \beta = \alpha \gamma n + \alpha \delta$$

or

$$\gamma \beta = \alpha \delta.\tag{3.44}$$

Inserting the values of  $\alpha$ ,  $\beta$ ,  $\gamma$ , and  $\delta$  from Eqs. (3.39) and (3.40),

$$-\frac{E_y}{V_T} \frac{J_{nx}}{q\mu_n V_T} = -\frac{E_x}{V_T} \frac{J_{ny}}{q\mu_n V_T}$$

or

$$\frac{E_y}{E_x} = \frac{J_{ny}}{J_{nx}}.\tag{3.45}$$

Therefore, a solution exists if and only if  $\vec{E}$  and  $\vec{J}$  are collinear; that is, only if the problem is one-dimensional.

### 3.3 Summary

The model equations were introduced and finite-difference discretization on an irregular triangular grid was described. Discretization of Poisson's equation on an acute triangle was shown and subtleties of area weighting

schemes were addressed with the perpendicular bisector method shown to be preferable. Modifications to the area weighting were shown to be necessary for obtuse triangles. Also, coupling coefficients within individual triangles were seen to be negative for the edge opposite an obtuse angle to achieve proper conservation of electric flux. However, it was demonstrated that the total coupling coefficient (the sum from two adjacent triangles) need never be negative as long as the triangle edge opposite the obtuse angle is not a solution region boundary or an interface between dissimilar materials.

Discretization of the electron continuity equation was described in a quasi-two-dimensional form using the Scharfetter-Gummel algorithm for discretization of the electron transport equation. Obtuse triangles were also shown to require special consideration in computing the coupling coefficients for the continuity equation discretization. The absence of a two-dimensional form of the Scharfetter-Gummel algorithm was demonstrated.

The next chapter covers the various methods of solving the combined Poisson and continuity discretized equations, factors affecting convergence, and means of accelerating the convergence.



## Chapter 4

### SOLUTION TECHNIQUES

Discretization of Poisson's equation and the continuity equation results in two systems of equations which must be solved in order to determine the potentials and carrier concentrations. There are two aspects to the solution of the equations—one is the solution of the matrix equation  $M\vec{x} = \vec{y}$  representing either the discretized Poisson or continuity equation by itself, the other is the consistent solution of the two coupled equations together. This chapter deals with both of these aspects.

#### 4.1 Matrix Equation Solution Methods

In this section, various grid solution method combinations are discussed for both Poisson's equation and the continuity equation. It is important to note that the coefficient matrices in these matrix equations are sparse [4.1] — that is, most of the matrix elements are zero. The matrix elements are the coupling coefficients between nodes in the grid, so only those elements corresponding to coupled nodes will be non-zero. This sparsity must be exploited to minimize the equation solution time.

The matrix equation solution methods may be divided into two classes: direct and iterative. In the direct solution techniques, a linear equation solution is achieved in a deterministic number of steps of the solution algorithm. The number of steps depends only on the algorithm chosen and the connectivity of the grid (i.e. how the nodes are interconnected) and is independent

of the values of the coupling coefficients or solution. Iterative matrix solution techniques have a deterministic number of steps for one pass (iteration) through the algorithm, but several or many iterations of the algorithm must be made in order to obtain an accurate solution, thus the total number of steps required is non-deterministic. In particular, the number of iterations depends on the values of the coupling coefficients, the values of the solution, and the accuracy of solution desired.

Typically, the number of operations (multiplies and adds) for a direct solution is significantly larger than for a single pass through an iterative solution. As the number of iterations increases, however, the total number of operations required for a solution using an iterative algorithm may surpass that required for a direct solution, thus there is a break-even point where one method becomes more efficient than the other. Unfortunately, locating the break-even point is more of an art than a science.

#### 4.1.1 Poisson's Equation

As mentioned in Chapter 3, the coefficient matrix for Poisson's equation is symmetric and positive definite. The symmetry may be exploited to reduce coefficient storage and operation count. The positive definite characteristic is required for convergence of some of the iterative techniques.

Of the direct solution methods,  $LU$  decomposition [4.2] is the most common. Based on Gaussian elimination, it may be used to solve full or sparse matrix equations. Sparse  $LU$  decomposition merely ignores the zero entries in the coefficient matrix and, in this sense, may be considered a "brute force" method. The coefficient matrix,  $M$ , is decomposed into the product of a lower triangular matrix  $L$  and an upper triangular matrix  $U$  such that

$M = LU$ . The matrix equation  $LU\vec{x} = \vec{y}$  is then easily solved by using forward substitution on  $L\vec{w} = \vec{y}$  and backward substitution on  $U\vec{x} = \vec{w}$ .

The advantages of this method are its flexibility and stability. It can be used on literally any grid, with any discretization, and with almost any non-singular coefficient values. Since the PISCES program was written in order to study grid, discretization, various device types, and various regions of operation, this flexibility and stability was extremely desirable. The disadvantages of  $LU$  decomposition are that it requires large amounts of storage and has a high operation count for large grids. The large storage arises from the fact that the  $L$  and  $U$  matrices both have more non-zero entries than the coefficient matrix  $M$ , a condition known as fill. This also influences the operation count which is on the order of  $N^2$ , (written as  $O(N^2)$ ), where  $N$  is the number of nodes in the grid.

The second direct solution method of interest is the Fast Fourier Transform (FFT) method as applied to the solution of Poisson's equation by Hoekney [4.3] for plasma physics computations. It is quite the opposite of  $LU$  decomposition in the sense that it has small storage requirements and a low operation count, but is very inflexible in terms of grid or device type. The FFT method demands grids similar to those of Figures 2.01b or 2.01d - that is, the horizontal grid spacing must be uniform along any given line and the number of nodes along that line should be a power of two. The method is based on the orthogonality of Fourier harmonics - if a function is an electrostatic potential solution of Poisson's equation for a given charge distribution, then each Fourier harmonic of the function is a solution of Poisson's equation for the corresponding harmonic of the charge distribution. The method works most efficiently if the Fourier analysis is performed in one dimension only which is normally the lateral direction for semiconductor devices.

Briefly, the method may be developed as follows. Poisson's equation in two dimensions with constant permittivity may be written as

$$\frac{\partial^2 \psi(x, y)}{\partial x^2} + \frac{\partial^2 \psi(x, y)}{\partial y^2} = -\frac{\rho(x, y)}{\epsilon}. \quad (4.1)$$

The potential and charge density may be represented by their Fourier expansions in the  $x$  direction as

$$\psi(x, y) = \sum_k \psi^k(y) e^{i2\pi kx/l} \quad (4.2a)$$

and

$$\rho(x, y) = \sum_k \rho^k(y) e^{i2\pi kx/l} \quad (4.2b)$$

where  $\psi^k(y)$  and  $\rho^k(y)$  are the amplitudes of the  $k^{th}$  harmonics, and  $l$  is the width of the solution region. Orthogonality principles dictate that each harmonic must independently satisfy Poisson's equation, thus

$$\frac{\partial^2 \psi^k(y)}{\partial y^2} - \left(\frac{2\pi k}{l}\right)^2 \psi^k(y) = -\frac{\rho^k(y)}{\epsilon}. \quad (4.3)$$

In the discrete problem there are only as many harmonics as there are nodes in the  $x$  direction along the grid line, so for an  $m$  by  $n$  grid the problem has been reduced from a two-dimensional matrix equation of order  $N$  (where  $N = mn$ ) to  $n$  one-dimensional matrix equations of order  $m$ . Since the matrix equations are one-dimensional, the coefficient matrix is tridiagonal and very efficient solution methods can be used. The total operation count is  $O(N \ln m)$ .

There are several disadvantages to the use of this method. One is the strict requirement for a rectangular grid with uniform horizontal grid spacing. Another is the requirement that permittivity be constant in the lateral direction, limiting its application to devices with planar material interfaces

and surfaces. The inclusion of electrodes within the device also causes a problem due to the non-linearity of the equation when carrier statistics are included. Finally, the method solves the Poisson's equation, but cannot be extended to the Newton-step form of the equation as described in Chapter 3. The linearization term resulting from the Boltzmann dependence of charge on potential re-introduces the spacial coupling of the equations which the FFT method eliminated. It is possible to use the FFT method on sub-regions of the device and then couple these regions together using an iterative technique. This makes it possible to simulate more complex device structures, but some of the advantages of a direct method are lost. A version of the FFT method is used in the CUPID program [4.4] developed by J. Frey.

Of the several iterative techniques for solving the matrix equations, the most commonly used is successive overrelaxation (SOR)[4.5] or a variant, successive block overrelaxation (SBOR) usually implemented as successive line overrelaxation (SLOR). These methods have been well analyzed in the literature [4.6, 4.7], so they will be discussed only briefly here. Theoretical analysis of the SOR and SBOR methods has concentrated on rectangular grids and much has been written on the choice of optimum relaxation parameter values. Although the SOR techniques are applicable to any type of grid, the optimum relaxation parameter value is not easily obtained for non-rectangular grids, so the convergence rate of the method suffers.

Theoretically, the operation count of SOR is  $O(N^{3/2})$  but this rate is rarely observed in practice since the number of iterations is sensitive to the accuracy of the initial guess of the solution and to the values of the coupling coefficients. The convergence rate depends on the eigenvalues of an iteration matrix derived from the coefficient matrix. A necessary and sufficient condition for convergence is that the largest eigenvalue must be less

than one. This condition is guaranteed if the coefficient matrix is positive definite (as it is for the discretization of Poisson's equation as discussed in Chapter 3.) A modified SOR method is used in the TWIST program [4.8].

In SOR, the solution is updated for each grid point independently and in succession based on current values of the solution at neighboring points. In SLOR, the solution for an entire line of points is computed simultaneously based on the values of the solution at points adjacent to this line. This has the advantage of providing a consistent solution to the equation along a line of points instead of at only one point. As a result, convergence is generally obtained with fewer iterations. The disadvantage is that one must solve a set of simultaneous equations for each line. The disadvantage is not great, however, since very efficient solution techniques exist for solving the tridiagonal matrix equations which result. Thus there is generally a net decrease in the amount of work required for convergence over simple SOR. Theoretically, the operation count is  $O(Nm)$  where  $m$  is the number of lines in the iteration.

Use of SLOR normally limits one to a rectangular grid since the iteration is done on a line by line basis. (There is nothing to prevent selection of serpentine "lines" in an irregular triangular grid; however, there is a penalty in the computational overhead required to access these lines and the convergence properties of such a method are unknown.) For IGFET simulations on a rectangular grid, the lines chosen are typically vertical lines normal to the gate since the potentials are more strongly varying in this direction than in the lateral direction. One is, in effect, solving several one-dimensional problems side-by-side and coupling them together with the iteration. When lateral potential variations become large, alternating between use of the horizontal and the vertical lines may provide some increase in convergence rate; however,

the increase in computational overhead offsets most of the advantage. SLOR is used in the GEMINI program [4.9].

Another iterative method commonly used in semiconductor device analysis is Stone's method [4.10] also known as the Strongly Implicit Procedure (SIP). It is one of several iterative procedures based on modification of  $LU$  decomposition to reduce the fill problem addressed earlier. In SIP, an  $L$  and  $U$  matrix are computed such that  $LU = M + E$  where both  $L$  and  $U$  have non-zero elements only where  $M$  has non-zero elements, thus there is no fill. The elements of  $L$  and  $U$  are very easily computed from the elements of  $M$ . The matrix  $E$  has the same non-zero structure as  $M$  except for two additional non-zero diagonals. The combination  $M + E$  is such that the iteration

$$(M + E)\bar{x}^{n+1} = (M + E)\bar{x}^n - (M\bar{x}^n - \bar{y}) \quad 4.4$$

converges rapidly to the proper solution where  $\bar{x}^n$  is the estimate of the solution vector at the  $n^{th}$  iteration.

In computing the elements of  $L$  and  $U$ , an iteration parameter between zero and one is used to accelerate the convergence of the algorithm. The choice of this parameter is critical since too large a value will result in divergence and too small a value will result in slow convergence. Typically, this parameter is cycled through a range of values in order to insure an adequate degree of stability while maintaining a good convergence rate. For problems of interest, choosing an optimum set of parameter values is very difficult so conservative estimates are made yielding sub-optimum convergence. The operation count for this method is  $O(N \ln N)$ .

The method was developed for solving equations on a two-dimensional rectangular grid; however, it is extendable to three-dimensional grids and to multi-variable problems. The principal requirement is that the coefficient

matrix have a highly regular banded structure. This limits the grids to rectangular, rectangular based triangular, or cubic structures. Actually, it is the highly regular connectivity of the grid that is important, so that the grids may be distorted as long as they retain their regular connectivity. The SIP method is used in the CADDET program [4.11].

Another method based on modification of  $LU$  decomposition is the Incomplete Cholesky—Conjugate Gradient (ICCG) method [4.12]. The conjugate gradient method [4.13] begins with an estimate of the solution vector and iteratively improves this estimate by converging to the exact solution along a set of orthogonal error vectors. Thus, with infinite precision arithmetic, the exact solution would be obtained in  $N$  iterations since  $N$  orthogonal vectors completely span the solution space. In fact, if the matrix  $M$  has only  $R$  distinct eigenvalues, then the algorithm must converge in only  $R$  iterations. Finite precision arithmetic will result in slower convergence, but since exact solutions are generally not required, sufficient accuracy will normally be achieved in the number of iterations described above. Unfortunately, the coefficient matrix of Poisson's equation typically has widely spread and uniformly distributed eigenvalues so that the full  $N$  iterations are required. The incomplete Cholesky decomposition algorithm, however, provides a method of transforming the coefficient matrix into an approximation of the identity matrix. The eigenvalues of this approximate identity matrix are highly degenerate so only a few are distinct. As a result, the conjugate gradient method converges very rapidly when used with the transformed coefficient matrix.

For a symmetric, positive definite matrix (as is the discretized Poisson coefficient matrix) an  $LL^T$  decomposition may be obtained instead of an  $LU$  decomposition where  $L^T$  is the transpose of  $L$ . This is the Cholesky decom-



position,  $LL^T = M$ . This decomposition results in the same fill problem as does  $LU$  decomposition; however, the incomplete Cholesky decomposition avoids this fill by simply zeroing out all element locations in  $L$  which are zero in  $M$  so that  $L$  retains the sparseness of  $M$ . Since the elements of  $L$  are computed recursively, the zeroed elements influence the computation of later elements. The result is a relation

$$LL^T = M + E \quad 4.5$$

much like that obtained for SIP. Pre- and post-multiplying by inverses yields

$$L^{-1}M(L^T)^{-1} = I - L^{-1}E(L^T)^{-1}. \quad 4.6$$

If the error matrix term on the right hand side is small, then the left hand side is approximately the identity matrix and may be used in the conjugate gradient algorithm for rapid convergence. Physically,  $M^{-1}$  (the Greens function) represents the coupling between a node and its neighbors. Ignoring fill in the  $L$  matrix is effectively neglecting the Greens function coupling to the distant (non-adjacent) neighbors of a node.

The number of operations per iteration of this method is comparable to that of SBOR or SIP; however, the rapid convergence (especially for stiff problems) often results in significantly less total work. The method does not require any particular matrix structure so that any type of rectangular or triangular grid may be used. The author knows of no device simulation program currently employing the ICCG method.

#### 4.1.2 Continuity Equation

The continuity equation, as discretized in Chapter 3, does not yield a symmetric or positive definite matrix. As a result, some of the matrix solution

algorithms discussed in conjunction with the solution of Poisson's equation cannot be applied to the solution of the continuity equation. Since no strong statement may be made concerning the eigenvalues of the continuity equation coefficient matrix, the SOR and SBOR methods may converge very slowly or even diverge for some conditions. The FFT method is also not applicable because the Scharfetter-Gummel discretization results in a collection of node-to-node differential equations for the transport equation rather than a single global differential equation.

*LU* decomposition, on the other hand, is suitable for use on the continuity equation and is used in the PISCES program. No instabilities have been observed in using this method. Both the SIP and ICCG methods also appear suitable; however, a modified form of the ICCG method is required since the coefficient matrix is not symmetric. For asymmetric matrices, the incomplete Cholesky decomposition must be replaced by an incomplete *LU* decomposition, where the incomplete *LU* decomposition algorithm is fully analogous to that for the incomplete Cholesky decomposition.

#### 4.1.3 Renumbering Algorithms

The total number of operations and data storage required for a matrix solution using *LU* decomposition is dependent not only on the number of nodes in the grid and how they are interconnected, but also on how the nodes are numbered. Consider, for example, the rectangular and rectangular-based triangular grids of Figures 2.01c and 2.01e. If the nodes in these grids are numbered from top to bottom along the left-hand column of  $m$  nodes and proceeding column by column from left to right along the  $n$  columns, then the structure of the coefficient matrices will appear as in Figures 4.1a and 4.1b. The bandwidths of the coefficient matrices are approximately  $2m$ .

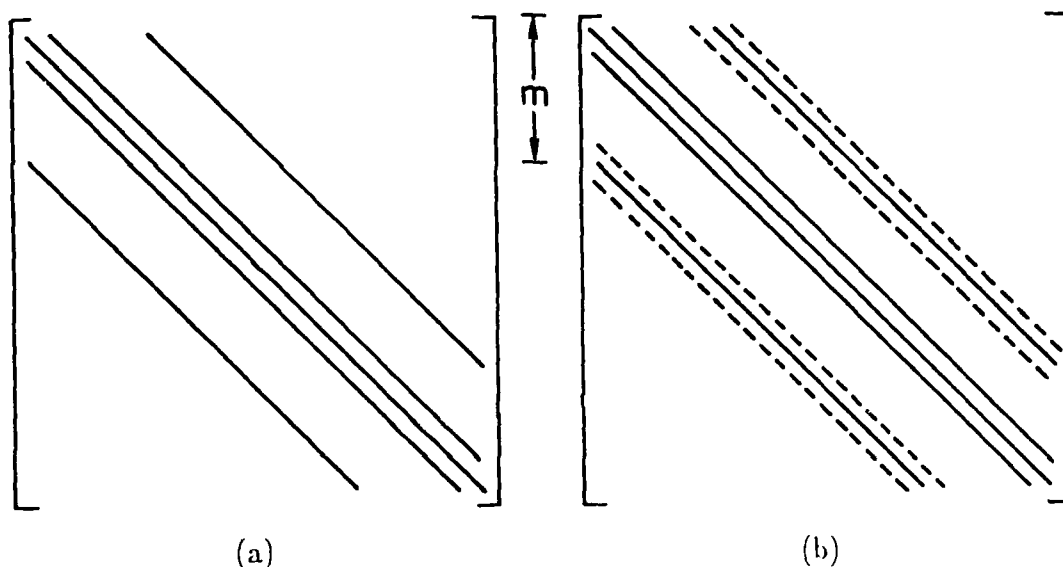


Fig. 4.1. Structure of coefficient matrix for: (a) rectangular grid, (b) rectangular based triangular grid. The dashed lines represent partially filled matrix diagonals resulting from the choice of the / or \ rectangle diagonals for subdivision into triangles.

Since the fill associated with  $LU$  decomposition will occur totally within the bandwidth of the matrix, the numbering of a rectangularly connected grid should proceed in the direction of the fewest nodes, by rows or by columns, in order to minimize the bandwidth.

Other more sophisticated grid numbering schemes exist including the method of nested dissection published by J. A. George [4.14]. Using this ordering, the operation count for  $LU$  decomposition may be reduced from  $O(N^2)$  to  $O(Nn)$  and the needed storage from  $O(Nn)$  to  $O(N \ln n)$ . An example of this ordering is shown for a 7 by 7 rectangular grid in Figure 4.2. Figure 4.2a shows the partitioning of the grid - all nodes labeled 1 are numbered first followed by those labeled 2 and then 3. Figure 4.2b shows the actual numbering with the resulting matrix structure shown in Figure 4.3.

1	2	1	3	1	2	1
2	2	2	3	2	2	2
1	2	1	3	1	2	1
3	3	3	3	3	3	3
1	2	1	3	1	2	1
2	2	2	3	2	2	2
1	2	1	3	1	2	1

(a)

1	19	5	44	9	29	13
17	20	18	45	27	30	28
2	21	6	46	10	31	14
37	38	39	40	41	42	43
3	24	7	47	11	34	15
22	25	23	48	32	35	33
4	26	8	49	12	36	16

(b)

Fig. 4.2. Nested dissection numbering scheme for a 7 by 7 grid. Part (a) shows the partitioning and part (b) the actual numbering of the grid.

Although the method was developed for rectangular grids, it provides similar benefits for the rectangular-based triangular grids used in PISCES. Using the same 7 by 7 grid and numbering as in Figure 4.2 but subdividing the grid into triangles, the matrix structure of Figure 4.4 results. The structure is essentially the same as that of Figure 4.3 and results in comparable improvements in operation count and storage.

The numbering scheme for nested dissection works perfectly only for square grids (*i.e.*  $m = n$ ) and only when  $m = 2^l - 1$  where  $l$  is an integer, such as 7 by 7, 15 by 15, or 31 by 31 node grids. For all other grids, slight perturbations in the numbering are required in order to achieve the same general pattern over the grid, thus there are many possible numbering patterns for any given grid. Several of these nested dissection numbering schemes were attempted for a 323 node (17 by 19) PISCES grid. Table 4.1 shows the number of non-zero entries in the  $L$  and  $U$  matrices for three of these numberings plus the normal column numbering.

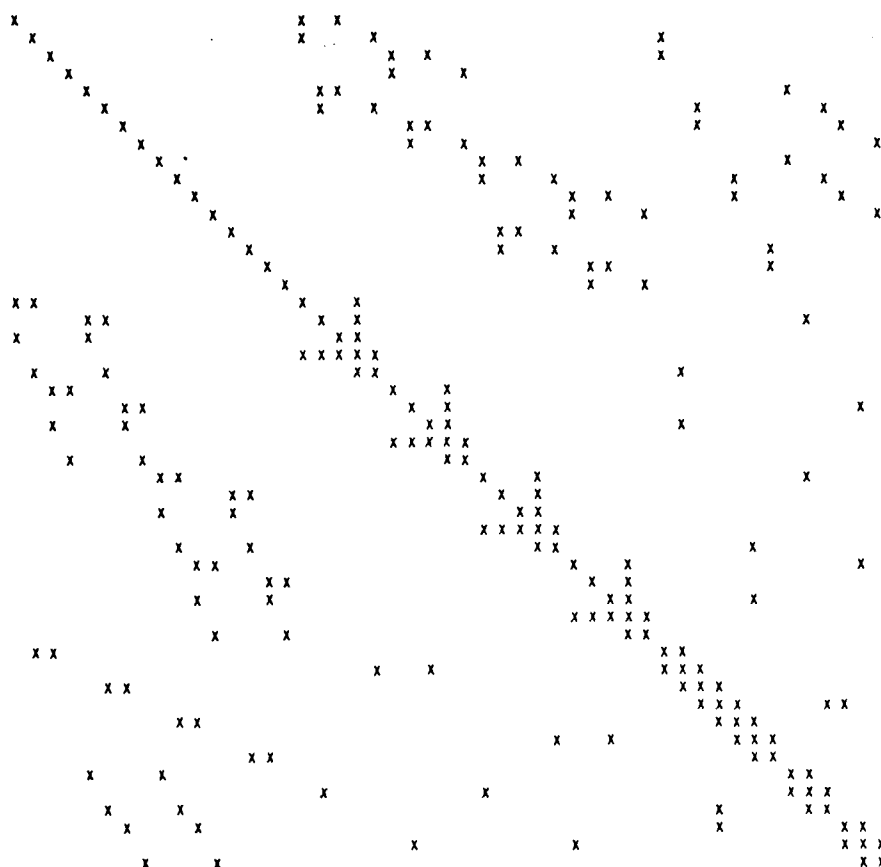


Fig. 4.3. Coefficient matrix structure for the nested dissection of Figure 4.2 on a rectangular grid.

The *symmetric* numbering is a perfectly symmetric dissection about the vertical and horizontal centerlines of the grid. The *automatic* numbering is roughly the same but not as symmetric. It is generated by an algorithm developed by I. S. Duff [4.15] and modified by Prof. R. J. Lomax of the University of Michigan. The *truncated* numbering is simply a corner of a perfect dissection; that is, beginning at the upper-left-hand corner node, the perfect dissection pattern is followed horizontally and vertically until running

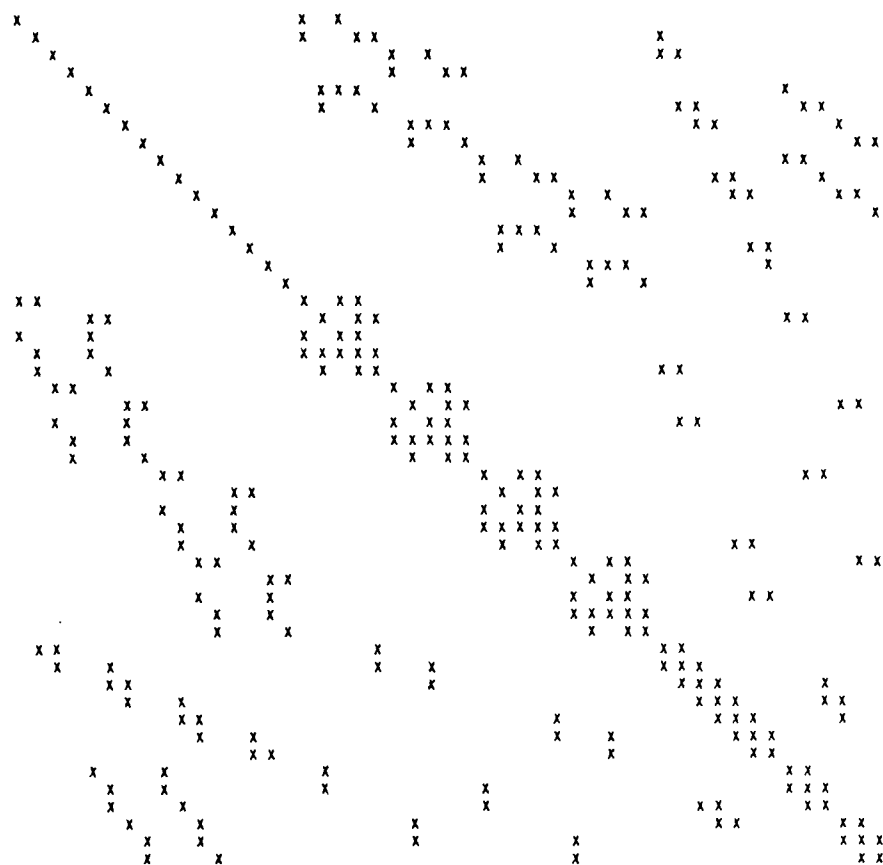


Fig. 4.4. Coefficient matrix structure for the nested dissection of Figure 4.2 on a rectangular based triangular grid.

out of grid. The results show that the symmetric numbering scheme has the greatest reduction in storage (20%) followed closely by the automatic numbering. The truncated numbering results in an increase in storage. Note that the original coefficient matrix contains only 2223 non-zero entries while the combined  $L$  and  $U$  matrices of the best numbering contain 6732. This difference is the fill generated in the decomposition process.

A decrease in storage implies a decrease in the number of operations

Table 4.1

## Nested Dissection Efficiencies

Numbering Scheme	L+U Storage (#)	Solution Time (sec)
Normal	8308	2.0
Symmetric N. D.	6732	1.8
Automatic N. D.	6948	1.9
Truncated N. D.	8554	2.9

323 Node grid with coefficient matrix  
storage = 2223 on DEC-10 computer.

required since fewer fill elements are computed. This is, in fact, the case as can be seen in the seconds-per-solution column of Table 4.1. Note that this table of data was obtained on a preliminary version of PISCES running on a large time-share computer so the execution times are somewhat biased by load variations. Additional timing comparisons were obtained on a 525 node (21 by 25) grid. These results are summarized in Table 4.2 and show that the improvements in storage and solution time generally agree with that predicted by theory and also that the improvements are greater for larger grids. It appears that the nested dissection method provides as much improvement for the rectangular based triangular grids of PISCES as for the rectangular grids for which it was developed.

There is one subtle drawback to the use of nested dissection for moderate grid sizes. Table 4.3 shows the average matrix solution time for PISCES on the IIP-1000F for three different conditions - scalar arithmetic, vector arithmetic and vector arithmetic with dissection. The PISCES program uses a

Table 4.2

## Nested Dissection Efficiencies versus Grid Size

Grid Size	L+U Storage			Solution Time (sec)		
	Normal	Dissected	Ratio	Normal	Dissected	Ratio
323	8308	6948	.84	2.0	1.9	.96
525	20924	12900	.62	5.8	3.9	.66
Ratio	2.5	1.9		2.9	2.0	
Theory	2.1	1.8		2.7	2.1	

matrix equation solution package entitled Vectorized General Sparsity algorithms (VEGES) [4.16] which puts all vectorizable steps in the *LU* decomposition process into vector operation form. On true vector architecture computers, significant time savings can be realized by exploiting the vectorizable operations; however, even on non-vector machines (such as the HP-1000F) some savings can be achieved with microcoded vector instructions. In Table 4.3, the *scalar* version of the program uses only scalar operations while the *vector* version uses the Vector Instruction Set of the HP-1000F computer. The vector version shows a factor of three improvement in solution time over the scalar version. When nested dissection is added to the vectorized version, however, the solution time increases. This increase is due to a reduction in the average vector length caused by the nested dissection renumbering. Depending on the computer, operations on short vectors can take longer than equivalent scalar operations due to the overhead associated with vector operation startup. The point of diminishing returns for vector versus scalar operation depends solely on the computer being used. As grid size increases,



Table 4.3

Nested Dissection Effect on Vector Arithmetic

Conditions	Solution Time (sec)
Scalar	18
Vector	6
Vector+ Dissection	8

HP-1000F Computer

however, nested dissection provides a net solution time reduction in spite of vector shortening. The PISCES program contains automatic nested dissection renumbering as a user controlled option.

Other renumbering schemes were also attempted including various forms of diagonal numbering and clustered numbering. None of these methods showed any significant reduction in storage; in fact, most showed significant increases.

#### 4.2 *Solution of Coupled Equations*

This section addresses possible ways in which the coupled equations, Poisson and continuity, can be solved so that each is consistent with the other. Factors affecting the degree of coupling, factors influencing the convergence rates of iterative methods and ways of accelerating convergence are covered.

AD-A119 110

AIR FORCE INST OF TECH WRIGHT-PATTERSON AFB OH F/6 9/1  
TWO-DIMENSIONAL NUMERICAL SIMULATION OF SEMICONDUCTOR DEVICES.(U)  
MAY 82 C H PRICE  
AFIT/CI/NR/82-240

UNCLASSIFIED

NL

202  
ADA  
19110

END  
DATE  
FILMED  
10-82  
DTIC

#### 4.2.1 Solution Methods

Figure 4.5 shows flow charts of the two principal approaches to solving the coupled equations—the simultaneous method and the alternating (or Gummels) method. The comparison between these two methods is somewhat analogous to the comparison of direct versus iterative matrix solution methods. The alternating method takes less work per pass but may require many passes and thus more work for a consistent solution than the simultaneous method.

The simultaneous method involves appending the two matrix equations together to form a single matrix equation which is twice as large.<sup>1</sup> In addition, the partial derivatives of all combinations of potential and carrier concentration for adjacent nodes must be included. This doubles the number of rows and approximately doubles the number of non-zero entries per row of the matrix since the carrier concentration at each node is coupled not only to the carrier concentrations of each adjacent node, but also to the potentials. The result is a matrix equation with nearly four times the number of non-zero entries as either coefficient matrix alone, and more than four times as many operations per solution. This rapid multiplication of effort is prohibitive for grids of any practical size. Additionally, the computation of the partial derivatives becomes more difficult as higher order physical phenomena such as field dependent mobility are introduced. For these reasons, most device simulation program designers have shunned the simultaneous method; however, some insist that simultaneous solutions are imperative for certain bias conditions [4.17].

---

<sup>1</sup>For two-carrier simulation, both continuity equation matrices would be appended resulting in a matrix three times as large.

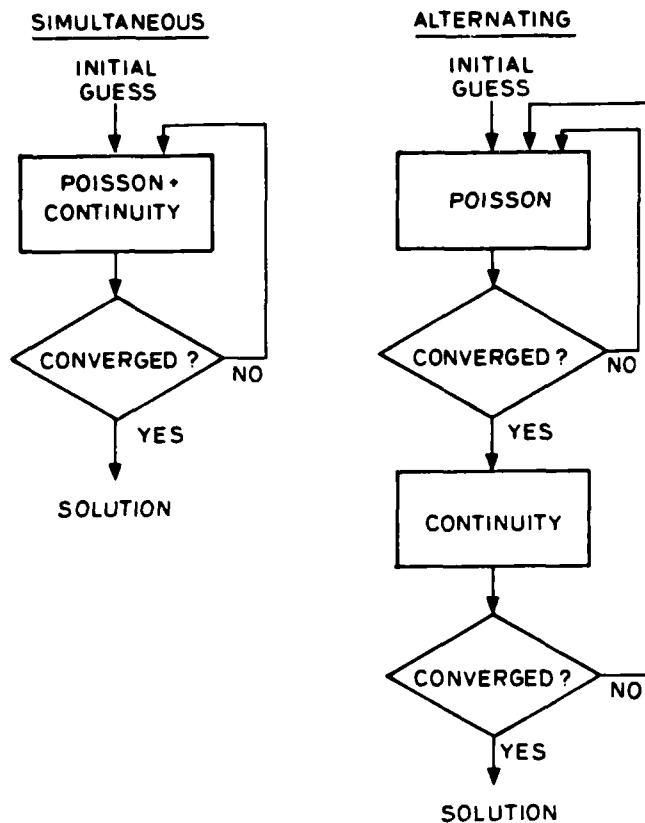


Fig. 4.5. Algorithm flow of simultaneous and alternating methods for solution of the coupled equations.

The alternating method of Gummel [4.18] is the most commonly used method for obtaining consistent solutions to the Poisson and continuity equations. As illustrated in the flow chart of Figure 4.5, beginning with an initial guess of potentials and carrier concentrations, the Poisson equation is solved for node potentials. The carrier concentrations are updated based on the new potentials and appropriate carrier statistics (Boltzmann or Fermi-Dirac) and the Poisson equation is solved again. This procedure is repeated until the potential change is below some convergence criteria limit. This is the inner

loop of the flow chart of Figure 4.5b.

Implicit in the updating of carrier concentrations based on potential changes and carrier statistics is the assumption of a fixed quasi-Fermi level. In fact, the quasi-Fermi level is an unknown which must be determined. This is done implicitly in the outer loop by solving the continuity equation and updating the carrier concentrations without changing the potentials. These new carrier concentrations require that the Poisson equation be solved again so the algorithm loops back to the top of the flow chart. Eventually, potentials and carrier concentrations converge and are consistent with both the Poisson and continuity equations.

The principal advantage of the alternating method is that less work is expended on each pass through the outer loop than in the simultaneous method. A significant disadvantage, however, is that the convergence rate of the method is dependent on the device operating conditions and may be very slow. In the simulations run on the PISCES program, for example, the number of outer loop iterations required for convergence without acceleration varied from one to sixty depending on the device biasing conditions. The reduction of this large number of iterations was a major thrust of the present work.

#### 4.2.2 Convergence Acceleration for the Alternating Method

Table 4.4 shows a typical convergence pattern for a MOSFET biased below threshold. The left hand column is the iteration count of the outer loop of the alternating method, the center column is the error measure for each iteration of the inner loop (Newton iteration on Poisson's equation), and the right hand column is the error measure for each iteration of the outer loop. The inner loop error measure is the average of the absolute values (the

Table 4.4

## Subthreshold Convergence

Outer Loop Count	Inner Loop Error	Outer Loop Error
1	3.7E-4	3.7E-4
	5.8E-12	
	1.5E-1	
	1.4E-2	
	1.7E-3	
	3.4E-4	
2	1.3E-5	1.4E-2
	3.3E-8	
	1.2E-12	
3		1.2E-12
Total	9	3

Total matrix solutions = 12  
 Gate = .5V, Drain = .01V

one norm of numerical analysis) of the incremental potentials resulting from the Poisson solution. The outer loop error measure is the one norm of the net change in the node potentials from one pass through the outer loop to the next.

Analyzing the convergence of Table 4.4, on the first pass through the outer loop, Poisson's equation converges in two iterations. The continuity equation solution, however, alters the carrier concentrations and results in six Poisson iterations on the second pass through the outer loop. On the third pass, only one Poisson solution is required and the change in potential is extremely small indicating that the algorithm has converged. Twelve matrix solutions are required for this simulation, nine for Poisson and three for continuity. This rapid convergence is due to the very weak coupling between

the Poisson and continuity equations. When a MOSFET is biased below threshold, the dominant charge (besides boundary charge) is the space charge of ionized impurities in the depletion regions. Since this charge is immobile, the continuity equation has little effect and a Poisson solution is essentially all that is required.

Another point worth noting in Table 4.4 is the convergence rate of the inner loop. Newton's method has quadratic convergence which means that when the solution estimate is in the vicinity of the correct solution, the error in the solution estimate will be squared with each iteration. This type of behavior is evident in the inner loop error measure shown in the table.

A different convergence pattern is seen in the MOSFET linear region simulation of Table 4.5. In this case, there is an inversion layer of free carriers at the semiconductor surface resulting in stiff coupling between the Poisson and continuity equations. Only pieces of the complete convergence sequence are shown, but it is clear that each continuity solution alters the carrier concentrations enough to negate the previous Poisson solution and cause several more iterations of the inner loop. A total of 151 inner loop and 46 outer loop iterations are required. It is this behavior which undermines the utility of the alternating method in device simulation. In the search for ways to reduce the simulation time required for devices above threshold, four techniques were derived which may be used to reduce this time by a factor of three to four. These techniques are detailed in the following paragraphs.

#### 4.2.2.1 Projection of the Initial Guess

The first convergence acceleration technique is the use of *projection* of the initial guess from previous solutions. In the PISCES program, the first bias condition solved for newly generated device grids is the flat band case.

Table 4.5

## Linear Region Convergence

Outer Loop Count	Inner Loop Error	Outer Loop Error
1	1.0E-2	6.9E-3
	2.8E-3	
	2.1E-3	
	6.1E-4	
	1.1E-4	
	4.2E-6	
	4.5E-2	
	2.6E-2	
	1.3E-2	
	2.3E-3	
2	1.1E-3	2.2E-2
	3.5E-4	
	4.3E-5	
.	.	.
.	.	.
.	.	.
16	.	2.8E-3
17	3.3E-3	2.5E-3
	6.6E-4	
	6.6E-5	
.	.	.
.	.	.
.	.	.
23	.	1.1E-3
24	1.0E-3	9.4E-4
	9.0E-5	
	.	
.	.	.
.	.	.
.	.	.
45	.	1.0E-4
46	8.3E-5	8.3E-5
	.	
Total	151	46

Total matrix solutions = 197  
 Gate = 2.7V, Drain = 1.0V



The initial guess for this case is the charge neutral condition. When not using the projection technique, the initial guess for every simulation thereafter is the previous solution. If a sequence of bias steps is being simulated, for example, each solution is used as the initial guess for the next bias. The result is slow initial convergence and occasional instability if too large a bias step is attempted.

The projection method involves extrapolating the initial guess for the new bias condition based on the solutions at two previous bias conditions. Only one contact bias may be varying between the two previous biases and the new bias. The assumption is that the potentials and quasi-Fermi levels for each node will vary linearly with the bias. Given two solution files for biases of  $V_1$  and  $V_2$  and a new bias of  $V_3$ , an extrapolation factor is defined as

$$\alpha = \frac{V_3 - V_2}{V_2 - V_1}. \quad (4.7)$$

The projected initial guess for potential and quasi-Fermi level at the new bias,  $\psi_3$  and  $\phi_3$ , is then computed for every node in the device by the relations

$$\psi_3 = \psi_2 + \alpha(\psi_2 - \psi_1) \quad (4.8)$$

and

$$\phi_3 = \phi_2 + \alpha(\phi_2 - \phi_1). \quad (4.9)$$

It is easily shown that these relations lead to an extrapolation of the electron concentration (assuming Boltzmann statistics) of

$$n_3 = n_2 \left( \frac{n_2}{n_1} \right)^\alpha. \quad (4.10)$$

These relations provide very good initial guesses. In charge neutral regions, the estimates are very accurate. In depletion regions the estimates

are not as good, but since there is little mobile charge there the coupling between Poisson's equation and the continuity equation is weak so that those errors which do exist are quickly corrected by Poisson solutions. Only at depletion edges and inversion layers do the estimates produce significant errors, but even there the estimates provide a smooth stable initial guess with excellent convergence probability. The combination of these properties serve not only to speed convergence but also to greatly increase the allowable bias step size. Another advantage of this method is its simplicity since all regions of a device are processed identically.

The projection method provides the greatest convergence acceleration when stepping the drain bias of devices in saturation, typically reducing the number of iterations by a factor of two. It is nearly as effective in the linear region but tends to lose some of its effectiveness in projecting initial guesses through the transition from linear to saturation. For subthreshold bias conditions the method does not significantly reduce the number of iterations; however it does allow larger bias steps to be taken without loss of stable convergence.

#### 4.2.2.2 Single Poisson's Equation Iteration

The second convergence acceleration technique is the use of a *single Poisson* solution per outer loop iteration. An inspection of Table 4.5 reveals that each continuity equation solution severely alters the previous Poisson solution as indicated by the large inner loop error seen after each outer loop iteration. In short, the accurate Poisson solution achieved through several iterations of the inner loop is unnecessary. By performing only one inner loop iteration on each pass through the outer loop, the number of outer loop iterations increases by roughly 20% but the total number of matrix solutions

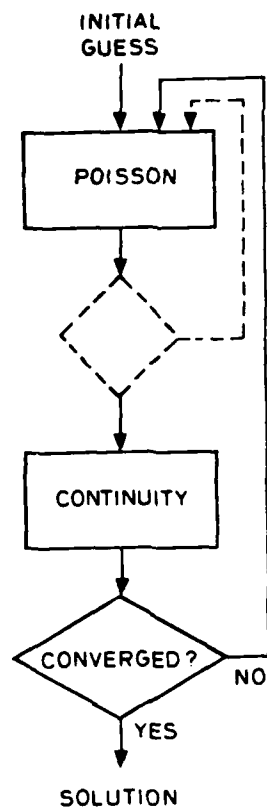


Fig. 4.6. Algorithm flow of the single Poisson acceleration scheme showing elimination of the inner loop.

decreases by roughly 40%. Figure 4.6 shows the revised flow chart with the elimination of the inner loop. This technique applies only to the linear and saturation bias conditions. Its use in subthreshold simulations will generally increase the total number of matrix solutions required.

The success and stability of this method provokes the question of whether this merging of the two iterative loops could be taken a step further when iterative methods are used for the matrix equation solutions. That is, rather than fully converging on an iterative solution to the Poisson matrix equation

and then fully converging on a solution to the continuity matrix equation, perhaps the program should alternate between a few iterations of each matrix solution. This possibility is worthy of additional consideration, but was not pursued in this work.

#### 4.2.2.3 Overrelaxation

The third convergence acceleration technique is the use of a form of *overrelaxation*. This technique must be used only with the single Poisson iteration described above. The method was developed by observing the details of the convergence of simulations using projection and a single Poisson solution. Figure 4.7 shows the potential and quasi-Fermi potential convergence of a node in the channel region of a MOSFET biased in saturation. The error is plotted versus the outer loop iteration count, thus each iteration represents two matrix solutions. Finer detail of the potential and quasi-Fermi potential convergence is shown in Figure 4.8 along with the electron concentration for the same node and the total drain current. Note that the latter two lag by about 20 iterations.

The monotonic nature of the potential convergence for this node over the first 10 to 20 iterations is typical of nodes observed in other regions of the device. Since the potential increments are nearly constant for every iteration, it appears that faster convergence can be obtained by merely increasing the size of the potential increments. This is somewhat analogous to overrelaxation in iterative matrix solution methods. When the vector of potential increments out of the Poisson solution is multiplied by a factor greater than one before being added to the previous potentials, faster convergence does indeed result.

Figure 4.9 shows the improvement in drain current convergence obtained using overrelaxation. A factor of 1.0 means no overrelaxation and 1.5 or 1.9

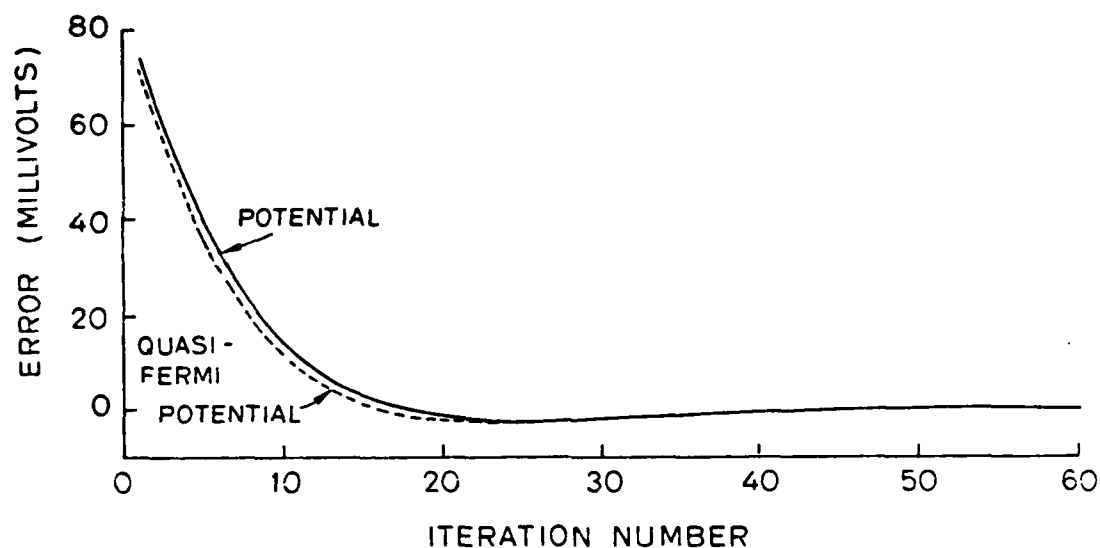


Fig. 4.7. Potential and quasi-Fermi potential convergence of a node in the channel region of a MOSFET biased in saturation.

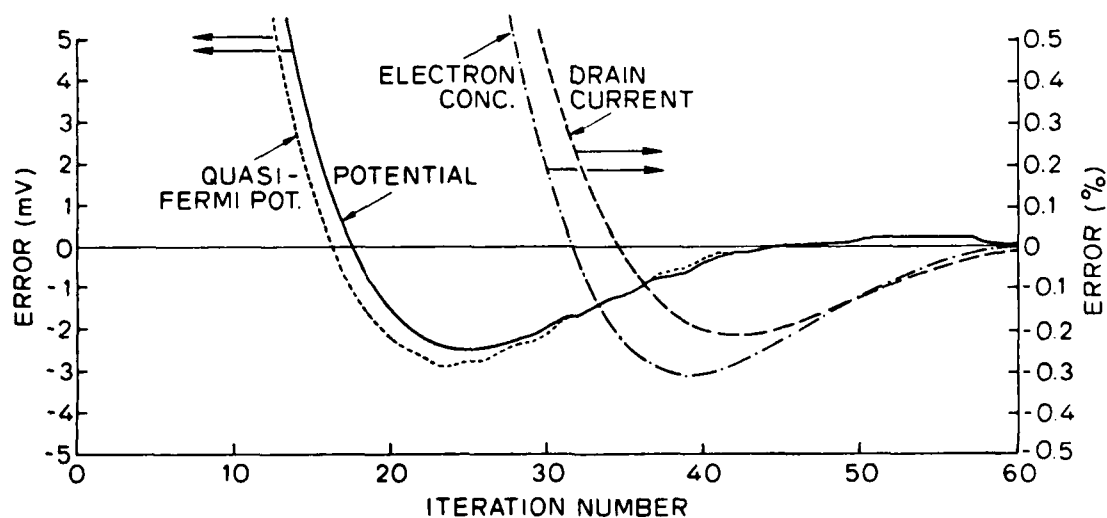


Fig. 4.8. Finer detail of the convergence shown in Figure 4.7 including node electron concentration convergence and total drain current convergence.

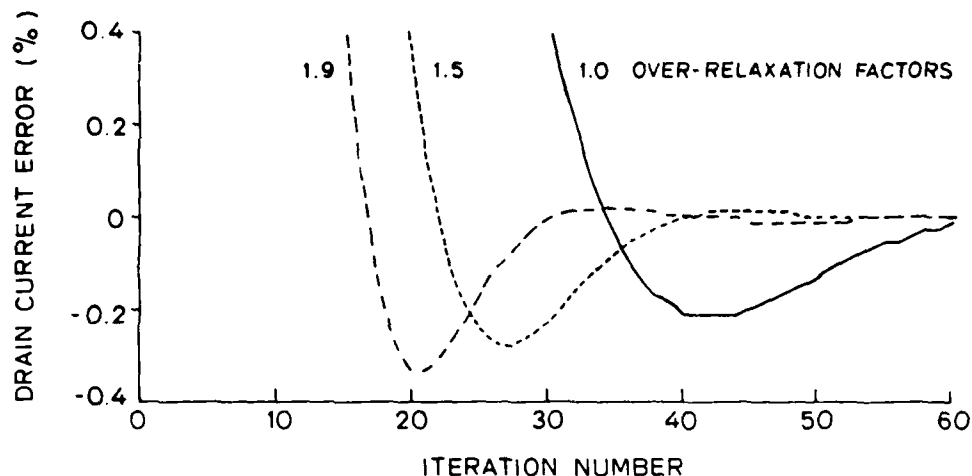


Fig. 4.9. Drain current convergence acceleration using overrelaxation.

mean that the Poisson solution vector is multiplied by 1.5 or 1.9 each time before updating the potentials. Factors of two or larger cause immediate instability in the iterations, suggesting a similarity between this method and the successive overrelaxation matrix iterative technique.

The PISCES program uses a variable overrelaxation factor which changes every three iterations, typically starting at a large value and then decreasing. This factor is computed using an algorithm published by Carré [4.19]; however, fixed factors between 1.5 and 1.9 generally work equally as well. The overrelaxation technique can be used only when using the single Poisson method also, thus it is applicable only to the linear and saturation regions of operation.

#### 4.2.2.4 Linearization Term Reduction

The final convergence acceleration technique is *reduction of the linearization term* in Poisson's equation. This method is closely related to the over-

relaxation method and their use is mutually exclusive. The basis for this method can be derived by inspecting Figure 4.7 and observing that the quasi-Fermi potential closely tracks the potential as they both converge. After each update of the potential by a Poisson solution, the continuity equation results in a nearly identical step in the quasi-Fermi potential. This violates one of the assumptions made in the discretization of Poisson's equation performed in Chapter 3. In that discretization, the linearization of the carrier statistics was based on the assumption that the quasi-Fermi level would remain constant as the potential changed resulting in a predictable change in electron concentration. When using the single Poisson method, however, the quasi-Fermi potential follows the potential change at each iteration; as a result the linearizing term in the discretization should be made smaller.

The Boltzmann carrier statistics assumed for this work are expressed for electrons as

$$n = n_i e^{(\psi - \phi_n)/V_T}. \quad (4.11)$$

The linearization of Poisson's equation then uses the partial derivative

$$\frac{\partial n}{\partial \psi} = \frac{n}{V_T}; \quad (4.12)$$

however, if  $\phi_n$  tracks  $\psi$ , then in one iteration of the outer loop,

$$\frac{\Delta n}{\Delta \psi} < \frac{n}{V_T}. \quad (4.13)$$

Thus, more rapid convergence should be obtained if one uses a linearization of

$$\frac{\partial n}{\partial \psi} = \alpha \frac{n}{V_T}, \quad (4.14)$$

where the accelerating factor,  $\alpha$ , has been empirically determined to lie in the range  $.3 < \alpha < .6$ .

This technique is roughly equivalent to the overrelaxation technique except that its effects are limited to regions of the device with high electron concentration. In low electron concentration regions, the linearization term is small and the Poisson solution values are unperturbed. In high electron concentration regions, the linearization term dominates and the size of the update potential solution is increased by a factor of approximately  $1/\alpha$ .

Figure 4.10 shows the drain current convergence acceleration achieved with this method. The normal unaccelerated case is obtained with a factor of 1.0. A factor of .5 is comparable to selective application of an overrelaxation factor of 2.0 and is seen to reduce the number of iterations by about one half. The third example shown used a factor which started at .2 for the first iteration, and increased by .033 each iteration until reaching .6 where it was fixed. The initial convergence was very rapid but unstable causing a residual ringing in the drain current even after the acceleration factor had moved into the stable range ( $\geq .5$ ) on the tenth iteration. The convergence rate in this case was only marginally better than that obtained using a fixed value of .5, but the rapid initial convergence obtainable from use of an unstable acceleration factor value will generally result in a savings of a few iterations. An acceleration factor starting at .3 and increasing by .04 each iteration until reaching .6 is recommended.

As mentioned earlier, this method may be used interchangeably with the overrelaxation method, but they cannot be used simultaneously. The performance of the two methods is nearly identical and no recommendations are made as to the use of one over the other.

It is important to note that Figures 4.9 and 4.10 show the drain current error within a range of plus or minus .4%. Drain current accuracies required for device simulation are more typically 1.0% and certainly not less than



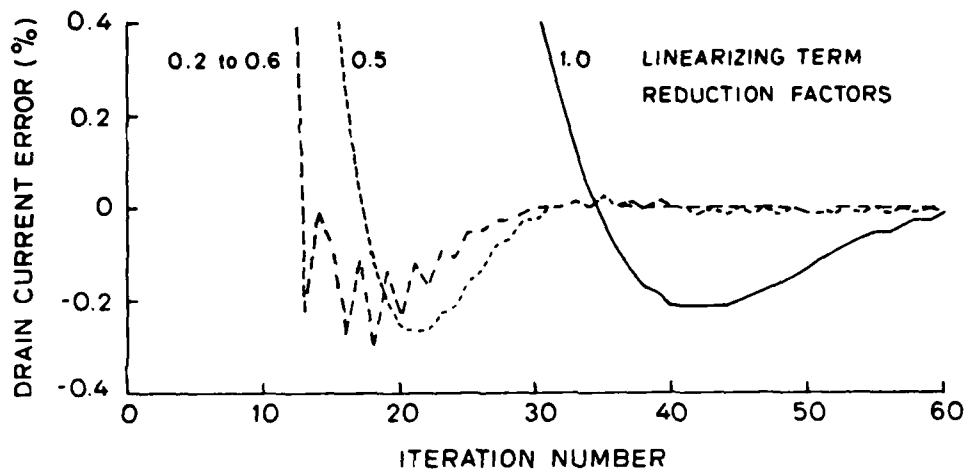


Fig. 4.10. Drain current convergence acceleration using reduction of the linearizing term.

0.1%. Looking at the figures, these levels of convergence are achieved in 12 and 25 iterations respectively, thus the convergence to useful levels is quite rapid when using the convergence acceleration techniques.

The convergence acceleration obtainable from combinations of these acceleration methods on a MOSFET operating in the saturation mode is exhibited in Table 4.6. The left hand column shows the five drain biases at which the simulations were run. The second column shows the number of matrix solutions (Poisson plus continuity) required for convergence at each drain bias with no acceleration. The total number required for all five biases is shown at the bottom along with a ratio of the total number for each column to the total of the unaccelerated column. The third column shows the acceleration achieved using projection alone. The fourth and fifth columns show the acceleration achieved using the single Poisson iterations alone and with projection. The final column shows the acceleration achieved by using all

Table 4.6  
Number of Matrix Solutions to Convergence

Drain Bias	Acceleration Method				
	None	Projection	1-Poisson	1-Poisson Projection	Overrelax. 1-Poisson Projection
1.5	189	98	120	74	68
2.0	188	110	86	84	68
2.5	183	103	88	84	48
3.0	183	78	88	70	58
3.5	184	82	90	52	48
Total	927	471	472	364	290
Ratio	1.00	.51	.51	.39	.31

three: projection, single Poisson, and overrelaxation. Comparable data was not taken for the linearizing term reduction method, but results are quite similar to the overrelaxation results.

The data in this table is slightly distorted in the conservative direction due to the convergence criteria which was being used at the time it was gathered. The iterations were stopped when the change in drain current from one iteration to the next became relatively small. This was determined later to be a poor criteria since it usually stopped on the peak of the drain current overshoot a local maximum of drain current error magnitude as seen in Figures 4.9 and 4.10. Occasionally, however, the algorithm missed the peak and stopped many iterations later resulting in a more tightly converged solution. Typically, the algorithm stopped on the peak for slower converging

methods but not on faster ones, thus the entries in the unaccelerated column of Table 4.6 are loosely converged while some of the entries in the remaining columns are tightly converged. As a result, the net improvement ratio is underestimated. The outer loop error norm as described earlier is a more stable measure of convergence and is used in the present version of PISCES. A convergence criteria for this error norm of  $10^{-4}$  results in loose convergence and  $10^{-5}$  in tight convergence. Use of this convergence criteria for all of the data in Table 4.6 would show a total acceleration improvement ratio of approximately .25.

#### 4.2.3 Convergence Rate Sensitivity to Bias Conditions

The previous discussions have referred to the sensitivity of convergence rate to the device region of operation. Figures 4.11 and 4.12 show this sensitivity in the computation of  $I_D/V_G$  and  $I_D/V_{DS}$  characteristics for a short channel IGFET using a 700 node grid. In both figures the drain current is shown as a solid line and the total solution time per bias point as dots. The solution time was measured on an HP-1000F minicomputer using vector arithmetic with each matrix solution of the 700 node grid requiring approximately 30 seconds. Note that the piecewise nature of the drain current curves results from the large discrete bias steps.

Figure 4.11 shows the drain current for a drain bias of .01V and a gate bias at steps of .2V from sub-threshold to the linear region of operation. Projection of the initial guess was the only convergence acceleration procedure used for these simulations. The very small drain bias in these simulations avoids the need for any other acceleration procedures. The solution times are small for all bias points but increase slightly near threshold where the device is changing from subthreshold to linear operating characteristics. This

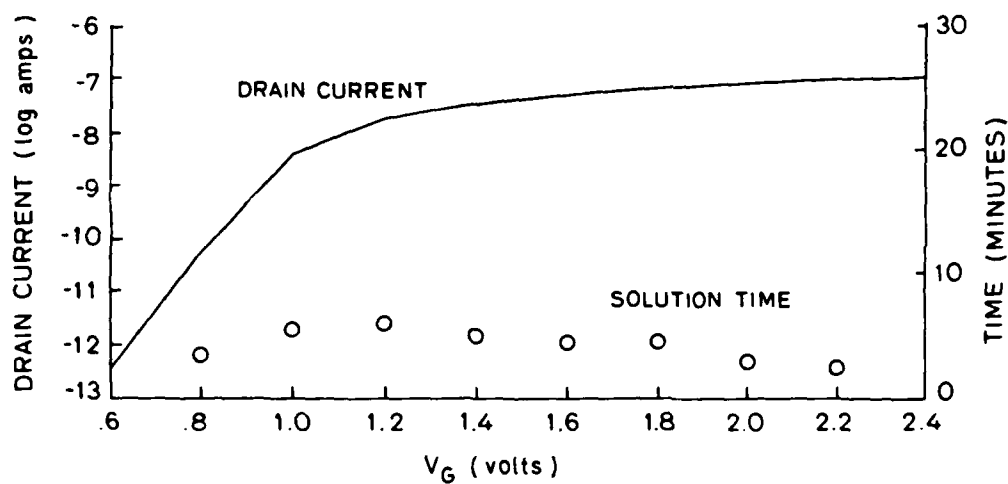


Fig. 4.11. Subthreshold and linear region simulation results at  $V_{DS} = .01V$  for .2V increments of  $V_G$ .

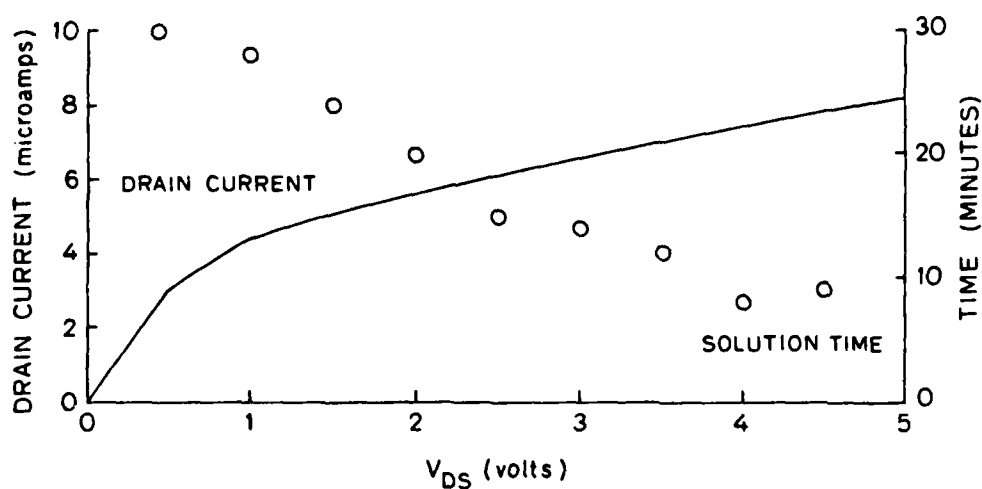


Fig. 4.12. Saturation region simulation results at  $V_G = 2V$  for .5V increments of  $V_{DS}$ .

increase is due to the errors incurred in projecting an initial guess in a region where the device is changing modes of operation.

Figure 4.12 shows data for the same device with a gate bias of 2V (well above threshold) and drain bias at .5V steps. The projection, single Poisson, and linearizing term reduction methods of convergence acceleration were used for these simulations. The noticeable aspects of this data are the large solution times required for drain biases at the knee of the curve and the decrease in solution time as the drain bias increases. The first data point shown on this figure is .5V and occurs in the transition region between the linear and saturation modes of operation of the device, thus the lower solution times required in the linear region are not shown. A small part of the reduction in solution times as drain bias increases is the improvement in accuracy of the projected initial guess, but the iteration output reveals that this accounts for only a few iterations difference per bias point. The principal difference is the rate of convergence. In the transition region it takes more than ten iterations to reduce the error norm by one order of magnitude, but in the saturation region it takes less than five. The cause of this performance is unknown; although, it appears to be a characteristic of the convergence acceleration schemes since it is not observed in unaccelerated simulations.

Some added insight is provided by Figure 4.13 which shows the error in surface potential at each of the first 18 iterations for a device with a  $1.5\ \mu\text{m}$  gate length. Only the projection and single Poisson acceleration schemes were used in this simulation. It is evident that the error is dominated by the first harmonic in spacial frequency between the source and drain and that it converges very slowly. Apparently the alternating method provides adequate local coupling (adjacent nodes) for the Poisson and continuity equations but is less adequate globally, specifically from the source to the drain along

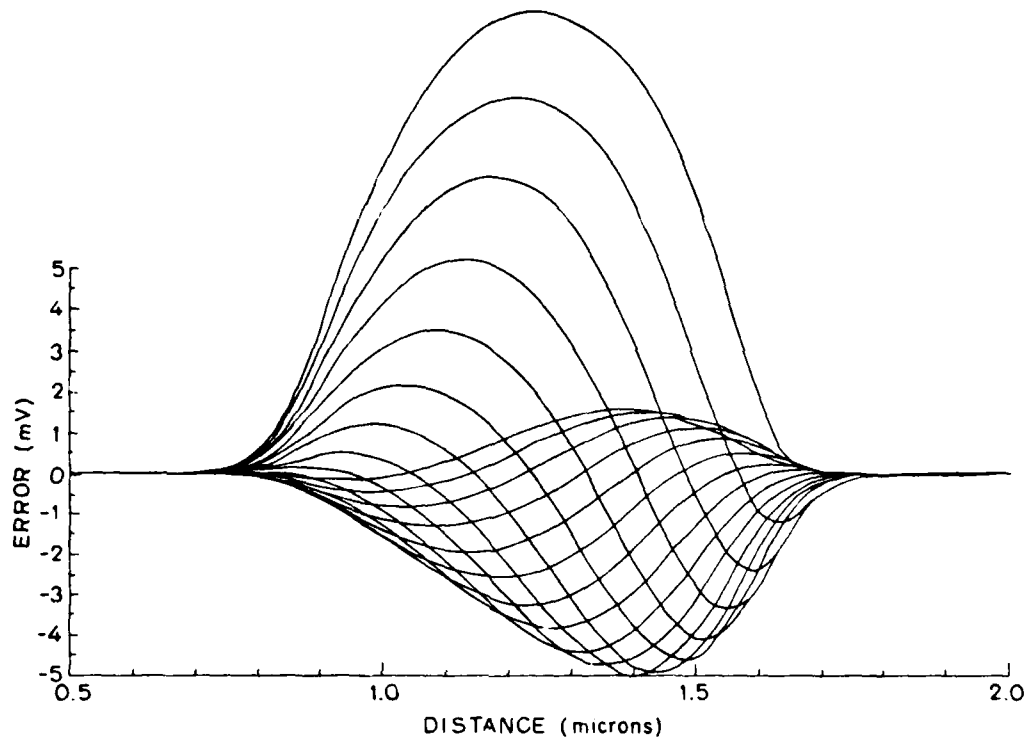


Fig. 4.13. Surface potential error at each iteration (of the first 18) for a MOSFET in saturation. The channel extends from approximately  $.75 \mu\text{m}$  to  $1.75 \mu\text{m}$ .

the inversion channel. It would appear prudent to investigate methods of supplementing this coupling across the length of the channel.

The approximate solution times for the various device operating regions are summarized in Table 4.7. An average number of matrix solutions required for convergence in each region is given along with the number of minutes required for three different grid sizes.

Table 4.7

## Approximate Solution Times on IIP-1000F

Operating Region	Number of Matrix Solutions	Minutes per Bias Point		
		323 Nodes	600 Nodes	700 Nodes
Subthreshold	7	1.2	2.7	3.5
Linear	25	4.2	9.6	12.5
Transition	50	8.3	19.1	25.0
Deep Saturation	15	2.5	5.8	7.5

#### 4.3 Summary

A variety of direct and iterative matrix solution methods are presented and their applicability to the Poisson and continuity equations are discussed. The advantages obtained from various grid renumbering schemes are also discussed. The nested dissection renumbering is shown to provide significant savings in both storage and solution time, especially for larger grids. These improvements are tempered by the fact that the re-ordering results in shorter vectors in the solution algorithm. The shorter vectors, in turn, slow down the processing speeds available on modern vector computers.

Solution of the coupled system of equations is addressed and the tradeoffs of simultaneous versus alternating solutions are discussed. Convergence of the alternating method without acceleration is analyzed and shown to be prohibitively slow for IGFETs biased in the linear and saturation modes. Four convergence acceleration techniques are presented which, in combination,

increase the convergence rate by roughly a factor of four. These methods involve computation of an improved initial guess, elimination of excessive solutions of Poisson's equation, overrelaxation of the potential updates, and reduction of the Poisson linearization term.

The convergence rate is examined as a function of bias for operation below threshold and operation above threshold with acceleration. Solutions above threshold take roughly two to six times as long as subthreshold solutions. Convergence is the slowest for devices biased in the transition region between linear and saturation operation but improves as the device is biased deeper into saturation. The dominant electrostatic potential error observed during convergence above threshold is a first harmonic in spacial frequency between the source and drain at the surface.

The next chapter provides two application examples of the PISCES program.



## Chapter 5

### APPLICATIONS

This chapter covers two examples of applications of the PISCES program. In the first example, an NMOS transistor with a channel implant is simulated in the punchthrough region of operation using three different simulation programs—PISCES, GEMINI [5.1], and CADDET [5.2]. The effects on punchthrough current of varying the source/drain junction depths are also shown. The second example demonstrates the utility of the PISCES program in evaluating physical models. The field dependent mobility of field effect devices is explored through use of a distance-from-the-surface mobility variation.

#### 5.1 MOSFET Punchthrough

A potential contour plot of the N-channel MOSFET simulated for this comparison is shown in Figure 5.1. The important device parameters are:

simulation length	$3.5 \mu\text{m}$
simulation depth	$5 \mu\text{m}$
oxide thickness	$500 \text{ \AA}$
channel length	$.7 \mu\text{m}$
substrate doping	$2 \times 10^{15} \text{ cm}^{-3}$
junction depth	$.4 \mu\text{m}$
channel implant dose	$3.4 \times 10^{11} \text{ cm}^{-2}$
channel implant depth	$.18 \mu\text{m}$

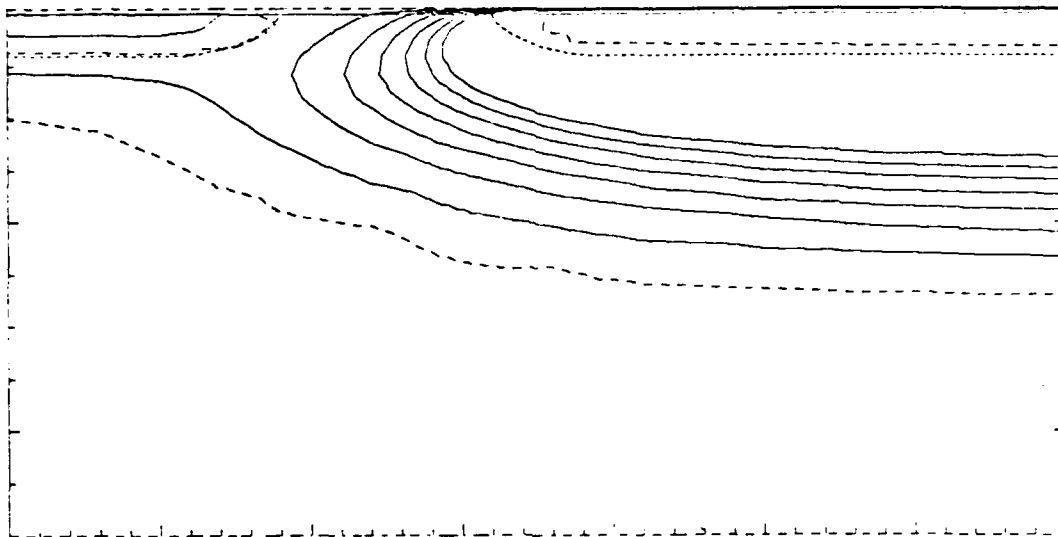


Fig. 5.1. Equipotential contour plot of N-channel MOSFET with implanted channel.

The planar oxide was used in order to accomodate the CADDET program comparison.

Figure 5.2 shows the results of the punchthrough simulations using PISCES, GEMINI, and CADDET. The drain bias was varied in one volt steps from one to ten volts while holding the gate at  $-5V$ . An analysis of equipotential contour plots would reveal that the punchthrough current path is at the surface for  $V_{DS} < 7V$  and in the bulk for  $V_{DS} > 7V$ . The  $.4\mu m$  source and drain regions are modeled as Gaussian implants with the concentration peak exactly at the semiconductor surface. The  $.45\mu m$  junction depth device source and drain have the same Gaussian shape but the peak is shifted  $.05\mu m$  below the semiconductor surface. Only PISCES was used to simulate the  $.45\mu m$  junction device. The punchthrough current for this device exceeds that of the  $.4\mu m$  device by as much as a factor of 50. This fifty-fold in-

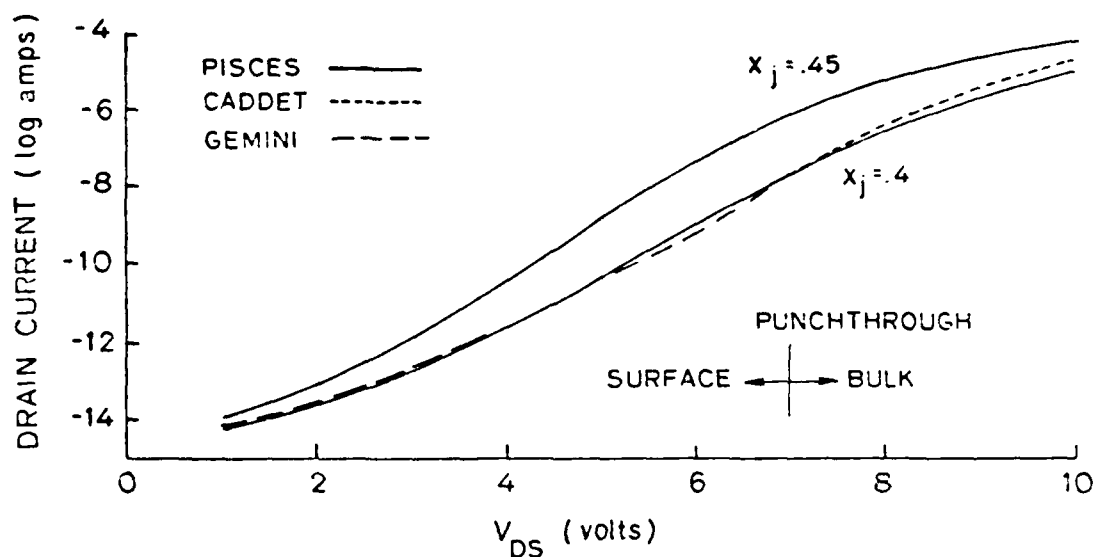


Fig. 5.2. MOSFET punchthrough characteristics using PISCES, GEMINI, and CADDET. The current path in the  $.4\mu\text{m}$  junction device is at the semiconductor surface for biases below 7V and in the bulk for higher biases. The  $.45\mu\text{m}$  junction device results are from PISCES only.

crease in punchthrough current with a 12% increase in junction depth is a rather startling result and illustrates the utility of numerical simulation in device design. This extreme sensitivity and similar sensitivities to junction curvature and impurity gradients make analytical modeling of these effects extremely difficult.

The three simulation programs used to analyze the  $.4\mu\text{m}$  junction depth device agree quite well throughout the range of drain biases. The reason for the discrepancy in the GEMINI solution at  $V_{DS} = 6V$  is unknown but may be related to the shifting of the dominant punchthrough current path from the surface to the bulk. The CADDET program shows a variation from the other two at drain voltages above 8V. The largest variation between the

three programs is approximately a factor of two. This is larger than one would like but certainly sufficient for the accuracies typically achieved for subthreshold device characteristics. Similar comparisons between PISCES and CADDET for devices operating above threshold show excellent agreement with variations typically less than 5%. Note that GEMINI is applicable only in the subthreshold and low-drain-bias linear regions of operation.

Inevitably, differences in results from the three programs can be traced to differences in grid placement, unrealistic physical assumptions for the simulated conditions, or loosely converged solutions. The automatic grid generation of CADDET and the inflexibility of the rectangular grid occasionally result in poorly placed grid points. The limited number of allowed grid points in PISCES also may result in sub-optimum grid placement. The CADDET assumption of a one-dimensional electric field in the oxide and the GEMINI assumption of constant quasi-Fermi levels can cause difficulty if the device structure or bias exceed valid ranges. Under-converged solutions are particularly noticeable in CADDET due to the solution methods and convergence criteria used.

All of these programs are more limited in accuracy by their models of higher order physical phenomena, however, than by their numerical methods. Accurate simulators must include models of phenomena such as velocity saturation, field and concentration dependent mobility, bandgap narrowing, degenerate statistics, surface states, two-dimensional impurity profiles, Schottky barriers, etc. The limitation is not one of implementation, but rather one of obtaining an accurate model. Much current controversy surrounds such topics as bandgap narrowing and mobility models. The utility of numerical simulation in investigating such models is displayed in the following section.

## 5.2 Strong Inversion Mobility

The subject of channel mobility under strong inversion has had considerable attention in recent years [5.3-5.5]. At issue is the variation of surface mobility with substrate doping, vertical electric field, crystal orientation, substrate bias, and interface charge.

Device designers have traditionally used empirical relations for predicting device performance based on the observed effects of substrate bias and impurity level on channel mobility. It was generally understood that some form of scattering caused the mobility variations but that the effects could be parameterized in terms of the substrate bias or impurity levels. More recently, it has become evident that the scattering occurs at the semiconductor-insulator interface and that the shape of the inversion layer charge profile is correlated with the mobility variations, *i.e.* the mobility reduction is greatest when the centroid of the inversion layer profile is nearest the surface. This compression of the inversion layer at the surface is related to the magnitude of the surface electric field, thus empirical models using the surface field as a parameter have emerged [5.6]. The observed variations with substrate doping and bias are caused by the difference in surface fields required for equivalent inversion levels at the different substrate impurity levels and biases. Even better agreement between model and measurement over a wide range of substrate conditions has been achieved with use of the "average" or effective field in the inversion layer [5.5].

From Gauss' law, the surface electric field is given by

$$E_S = \frac{Q_I + Q_B}{\epsilon_s} \quad (5.1)$$

while the effective field is given by

$$E_{eff} = \frac{Q_I/2 + Q_B}{\epsilon_s} \quad (5.2)$$

where  $Q_I$  is the inversion layer charge per unit area,  $Q_B$  is the bulk depletion charge per unit area, and  $\epsilon_s$  is the semiconductor permittivity. The current belief is that phonon and surface roughness scattering mechanisms dominate the room temperature surface mobility in strong inversion and phonon and coulombic (interface and oxide charge) scattering in weak inversion; however, the quantification of these effects is not well understood.

The point of the foregoing discussion is that the empirical and analytical models for surface mobility used by device designers and in circuit simulation programs necessarily avoid handling the basic underlying physical mechanisms in order to obtain computationally manageable models. Numerical simulation, on the other hand, can more easily accommodate the underlying physics and thus can be used by device designers for more accurate and operating-region-independent results, and by device physicists for studying the validity of their models. In this section, PISCES is used to evaluate a surface mobility model in which the mobility varies with distance from the surface.

A very long channel device was chosen for this study in order to minimize the influence of the source and drain on the channel potential. The important device parameters are:

oxide thickness	1000 Å
channel length	50 $\mu\text{m}$
substrate doping	$1.2 \times 10^{15} \text{ cm}^{-3}$
fixed interface charge	$10^{11} \text{ cm}^{-2}$ .

It is well known that the carrier mobility varies from its value in the bulk to a lower value at or near the surface; however, the value of the surface

mobility and the functional form of the variation with distance are not known. A. Gnädinger and H. Tally [5.7] have computed the thickness of inversion layers to be on the order of  $100 \text{ \AA}$  and have shown quantum mechanically that the carrier density peaks at distances on the order of  $25 \text{ \AA}$  from the surface. It is reasonable to assume that the variation with distance is monotonic, thus the mobility model chosen is one in which the mobility decreases exponentially with distance from the surface with characteristic length on the order of  $25 \text{ \AA}$ . Mathematically, it is expressed as

$$\mu(y) = \mu_b - (\mu_b - \mu_s)e^{-y/\sigma} \quad (5.3)$$

where  $\mu_b$  is the bulk mobility,  $\mu_s$  is the mobility exactly at the surface,  $y$  is the distance from the surface, and  $\sigma$  is the characteristic length of the variation. Figure 5.3 shows the results of simulations with three different values of the model parameters. In (a) there is no mobility variation with depth and the classical straight line variation of  $I_D$  with  $V_G$  is observed. In (b) the mobility is  $1000 \text{ cm}^2/\text{V-s}$  in the bulk,  $400 \text{ cm}^2/\text{V-s}$  at the surface and has a characteristic length of  $50 \text{ \AA}$ . In (c) the bulk mobility remains the same but the surface mobility is reduced to  $10 \text{ cm}^2/\text{V-s}$  and the characteristic length to  $33 \text{ \AA}$ . In order to accurately quantize these variations, the grid spacing perpendicular to the surface is very small, starting at less than  $10 \text{ \AA}$ . This small spacing is somewhat restrictive since it consumes large numbers of grid points and could cause numerical errors in the difference equations.

The reduced drain current for cases (b) and (c) of Figure 5.3 reflects the reduced channel mobility. The flattening out of the curves at higher gate biases is characteristic of IGFET's and results from the crowding of the inversion layer charge closer to the surface. Figure 5.4 shows the relation between the  $I_D/V_G$  curves, the field effect mobility ( $\mu_{FE}$ ), and effective

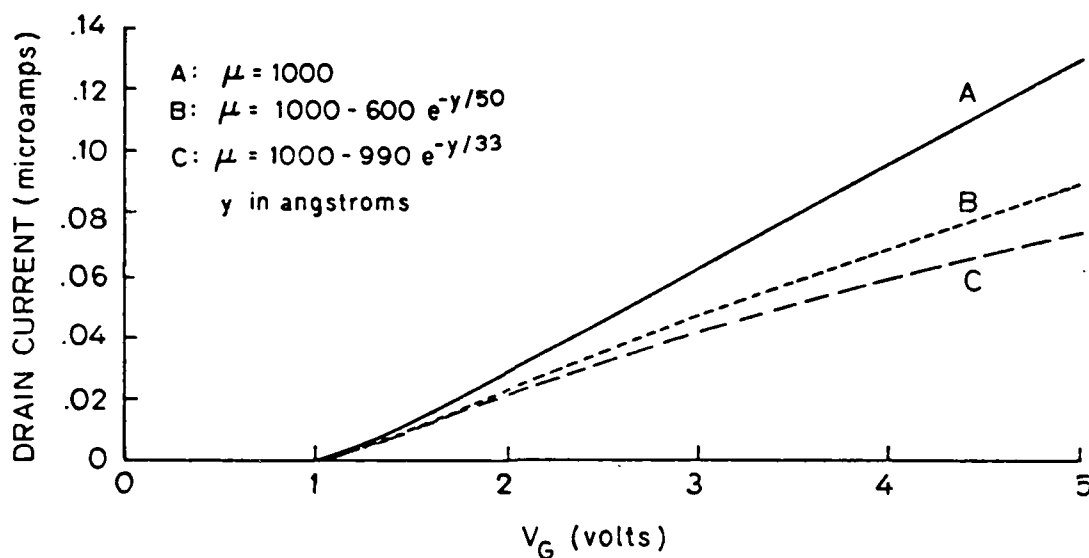


Fig. 5.3. Rolloff of drain current with gate bias showing effect of distance-from-the-surface mobility model. The mobility is in  $\text{cm}^2/\text{V-s}$  and  $y$  is in  $\text{\AA}$ .

mobility ( $\mu_{eff}$ ). These are expressed mathematically as

$$\mu_{eff} = \lim_{V_{DS} \rightarrow 0} \frac{(L/W)g_d}{qN_{inv}} \quad (5.4)$$

and

$$\mu_{FE} = \lim_{V_{DS} \rightarrow 0} \frac{(L/W)g_m}{C_0 V_{DS}} \quad (5.5)$$

where  $L$  and  $W$  are the channel length and width,  $g_d$  and  $g_m$  are the drain conductance and transconductance,  $qN_{inv}$  is the total induced charge in the channel per unit area, and  $C_0$  is the gate capacitance per unit area. Note from the figure that these two mobilities are equal at the steepest part of the curve. These relations are described by S.C. Sun [5.8] whose measurements are used for the comparisons in the remainder of this section.



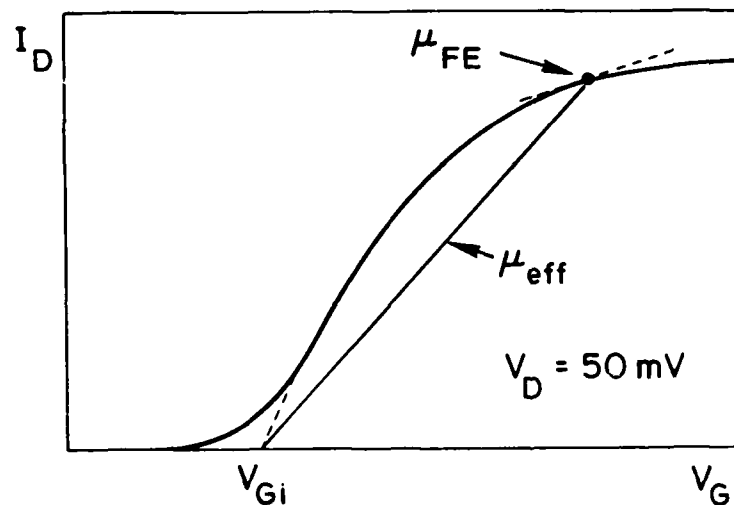


Fig. 5.4. Definition and relationship between effective mobility  $\mu_{eff}$  and field effect mobility  $\mu_{FE}$ .

Figure 5.5 shows the comparison of simulation versus measured results for the device described earlier. The mobility model parameters used are  $\mu_b = 1286 \text{ cm}^2/\text{V-s}$ ,  $\mu_s = 400 \text{ cm}^2/\text{V-s}$ , and  $\sigma = 50 \text{ \AA}$ . The bulk mobility value is chosen based on the substrate impurity concentration. The solid and dashed lines represent the simulated data for  $\mu_{eff}$  and  $\mu_{FE}$  respectively and the circles and squares the measured values. Although the simulated and measured results are offset, they have the same shape. Obviously, the distance-from-the-surface mobility model has the proper effect on the effective mobility although the functional form or parameter values of the model may not be correct. It would be of interest to fit the two unknown parameters,  $\mu_s$  and  $\sigma$ , to additional measured results in order to determine the parameter sensitivities to device fabrication and structure variations. The curve fitting would be an empirical study; however, the parameters have a physical basis

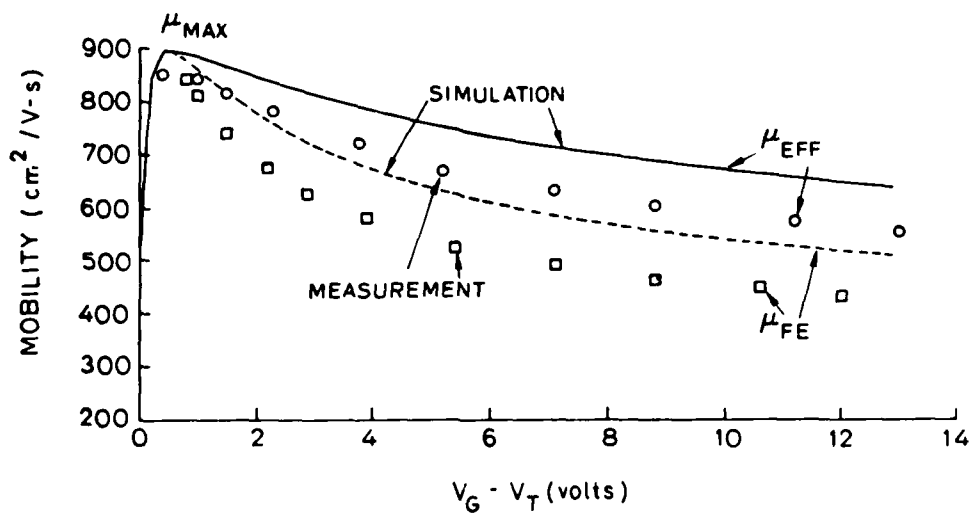


Fig. 5.5. Comparison of simulated and measured values of  $\mu_{eff}$  and  $\mu_{FE}$  with varying gate voltage.

and could shed light on mobility phenomena as opposed to analytical models such as those used in circuit simulation program device models which have no physical basis. This study is not attempted since it is not the intent of this work to develop a new mobility model, but it is suggested as an area of further research.

The maximum value of mobility occurs at the point labelled  $\mu_{max}$  in Figure 5.5. Since this peak always occurs at a gate voltage just above threshold, the surface is only weakly inverted. As substrate doping increases, the surface field required for the same degree of surface inversion also increases resulting in more crowding of the inversion layer charge and thus lower  $\mu_{max}$ . This reduction in maximum mobility with increased substrate doping is seen in Figure 5.6. Both measured and simulated results are shown along with the bulk mobility values for comparison. The simulated results were ob-

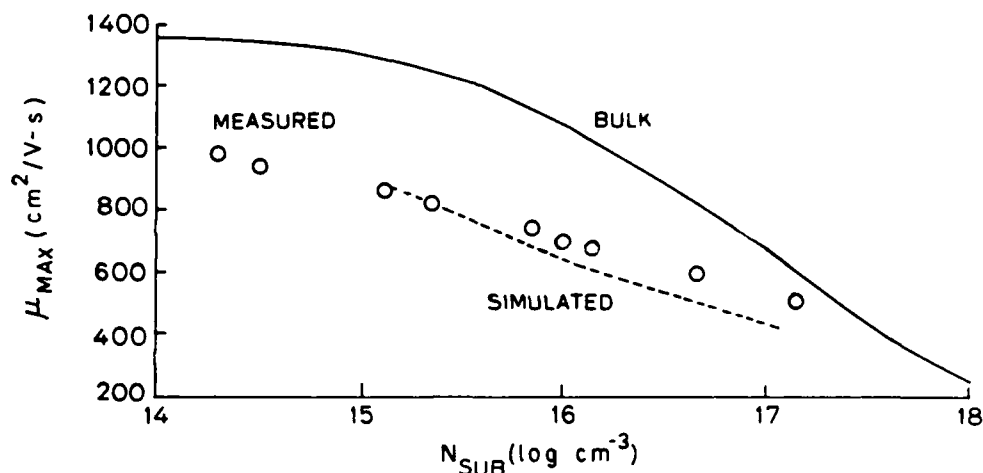


Fig. 5.6. Comparison of simulated and measured values of  $\mu_{max}$  sensitivity to substrate doping.

tained at substrate dopings of  $1.2 \times 10^{15}$ ,  $1.2 \times 10^{16}$ , and  $1.2 \times 10^{17} \text{ cm}^{-3}$ . The simulated results are comparable to the measured values but show a steeper slope indicating that the surface mobility reduction effect is too strong with the chosen parameters.

One of the principal points of this section is that numerical simulation can be useful in the evaluation of physical models. Attempting to match the measured data shown here and available elsewhere by modifying the mobility model should provide additional insight into the characteristics of surface inversion layer mobility. The other point is that the use of the most fundamental physical models possible results in the most powerful and versatile device simulation programs. The relatively simple distance-from-the-surface mobility model, for example, is easier to implement than an effective field mobility reduction model and much more universal than substrate bias and substrate doping mobility reduction models.

### 5.3 Summary

Two examples of application of the PISCES program are presented—punchthrough current simulations on an implanted channel NMOS transistor and evaluation of a distance-from-the-surface mobility model. The PISCES, GEMINI, and CADDET programs are compared for the punchthrough simulations. Their results agree throughout the range of simulated biases. Simulations also show that a 12% increase in source/drain junction depth can result in a 50 fold increase in punchthrough current. Simulations using the depth dependent mobility model indicate that such a model may be useful for device simulation in lieu of field dependent models. An exact form for the model is not pursued but is suggested as an area of further research. The effects of mobility reduction with increased gate bias and increased substrate doping are demonstrated. One disadvantage of the depth dependent mobility model is the need for very fine grid spacing normal to the surface.

The next chapter summarizes the conclusions and recommendations of this work.

## Chapter 6

### CONCLUSION

The two-dimensional structure of modern semiconductor devices demands the use of two-dimensional numerical simulation in device design. The application of analytical models simply cannot accurately account for the highly two-dimensional impurity profiles and potentials and their interaction.

All of the two-dimensional device simulation programs currently available are too restrictive to be of great service to device designers as witnessed by the slow acceptance of the few commercially available device simulation programs. Restrictions in allowable device structures, grid, computation time, memory requirements, accuracy of physical models, and ease of user interface are seen to some degree in all programs. Various aspects of these limitations have been addressed in this work through the development and application of a device simulation program, PISCES. The program contains some of the same limitations but is extremely flexible and allows investigation of many of the device simulation program restrictions.

#### *6.1 Summary*

The allocation of grid in various regions of a device has been addressed in terms of grid type, grid density, device structures, and impurity profiles. These analyses show that the densest grid is required in regions of high net charge density, large gradients of net charge density or large gradients of potential. The use of reflecting boundary conditions along the sides of the

device is shown to require significant lateral extensions of the source and drain in order to accurately represent the device potentials. Simulations have also demonstrated the inadequacy of rectangular uniformly doped approximations to the source/drain regions. Even a very coarse approximation to a Gaussian source/drain profile is shown to provide very good results.

The application of finite difference discretization of Poisson's equation and the current continuity equation to an irregular triangular grid has been presented including the special cases of obtuse triangles. A more consistent area allocation scheme has been presented along with a simple technique for avoiding negative coupling coefficients in the Poisson discretization for obtuse triangles. The quasi-two-dimensional discretization of the continuity equation using the Scharfetter-Gummel algorithm is accompanied by a proof of the non-existence of a fully two-dimensional form.

A variety of matrix solution techniques have been compared in terms of their applicability to device simulation. The flexibility and stability of the *LU* decomposition method are offset by the rapid growth of solution time and memory requirements with grid size. The opposite can be said for the iterative methods of SIP and ICCG which do not grow as rapidly with grid size but are more restrictive in grid type and sensitive to the matrix coefficients in solution time.

Evaluation of the *LU* decomposition method has shown that proper numbering of the grid can result in time and memory savings. Numbering a rectangularly connected grid in the shortest direction (*i.e.* row or column) minimizes the generation of non-zero matrix elements and thus reduces the matrix equation solution time. A more exotic renumbering scheme, nested dissection, is shown to reduce the solution time from  $O(N^2)$  to  $O(N^{3/2})$  and storage requirements from  $O(N^{3/2})$  to  $O(N \ln n)$ . These results, obtained on

triangular grids, match those reported for rectangular grids. A drawback of the nested dissection method is that vector computers cannot be used to full advantage because of the shorter vectors generated by this numbering.

The convergence properties of Gummel's alternating method for the solution of Poisson's equation and the continuity equation has been thoroughly analyzed. The method works exceedingly well for devices biased below threshold, generally requiring less than five iterations. Simulations of devices biased above threshold but with very little current flowing also converge rapidly. Significant convergence problems occur, however, for devices biased in the linear and saturation regions of operation where substantial currents are flowing.

Four different methods have been derived in the course of this work for accelerating the convergence of the alternating method in these operating regions-- projection of the initial guess from previous solutions, use of only one Poisson iteration per alternating iteration, overrelaxation of the electrostatic potential solutions and reduction of the Poisson linearizing term. The use of combinations of these techniques reduces the average solution time by approximately a factor of four. Using these acceleration techniques, as the drain bias increases the solution time for MOSFET drain bias steps is seen to increase until reaching saturation and then decrease as the device is biased deeper into saturation. This solution time reduction with increasing bias in the saturation region is seen only when using these acceleration techniques and tends to counter claims that the alternating method is not practical for simulation of devices above threshold.

Even with the acceleration techniques, the convergence above threshold is still relatively slow. Plots of surface potential error versus location along the channel show a single sinusoidal variation of the error from source to

drain. The sinusoid undergoes very slow decaying oscillations in a wave-like manner as the iterations progress. This first harmonic in spatial frequency of the surface potential error appears to be a dominant factor in the slow convergence of the alternating method above threshold.

Comparisons of the PISCES, GEMINI and CADDET device simulation programs has shown agreement to within a factor of two in subthreshold simulations and to within a few percent above threshold. These and other differences between simulation program solutions can usually be traced to difference in grid placement, physical assumptions or convergence tolerance.

A depth dependent mobility model has been implemented in order to examine its feasibility and to demonstrate the ability of numerical simulation programs to use more fundamental physical models than the analytical or empirical models. Simulations using this model have shown reasonable agreement with measurement for mobility variations with gate bias and substrate doping. This application of device simulation also demonstrates the utility of such programs in the evaluation of physical models.

## *6.2 Recommendations*

As with most scientific endeavors, there appear to be more questions at the end than there were at the start. The PISCES program is extremely well suited to the studies performed in this work but is unsatisfactory as a device designers tool. Several modifications to the program are suggested with varying degrees of additional research required.

The rectangularly connected triangular grid of PISCES appears to be a reasonable compromise in terms of flexibility in matching device structures,



ease of generation, and compatibility with matrix equation solution techniques. An automatic grid generation scheme should be developed for implementing this grid based on the grid density criteria described in this work. Grid refinement with changing bias should be studied with a consideration for the tradeoff between the increased accuracy achieved and the possibility of slight discontinuities in the device characteristics due to the changing grid.

The *LU* decomposition matrix solution method used in PISCES is not the optimum method for device simulation and should be replaced. Studies should be performed on comparisons of the SIP and ICCG methods as they appear to be the best suited replacements. Implementation of either of these iterative matrix solution techniques should be accompanied by study of the effectiveness of fully merging the alternating solution iterations by mixing the Poisson and continuity equation matrix iterations.

Although the obtuse triangle discretization is relatively sound and many apparently successful simulations have been performed with obtuse triangles in the grid, their effect is not fully understood. Some high resolution contour plots in regions with large numbers of highly obtuse triangles, for example, have shown slight distortions in the equipotential contours. Additional research is suggested to quantify these effects.

The slow convergence of the alternating method for simulation of devices above threshold remains a problem in spite of the acceleration achieved in this work. The significance of the oscillations seen in the surface potential should be investigated with the possibility of supplementing the Poisson and continuity equation coupling along the length of the channel.

A more definitive study of the depth dependent mobility model should be performed in order to determine the proper functional form and parameter values for the model and its practicality for use in device simulation. The

grid density requirements for implementation of the model should also be considered. Other poorly understood phenomena such as weak inversion mobility should also be examined via device simulation.

## Appendix A

### PISCES DEMONSTRATION EXAMPLE

Parts of the PISCES program have been described in the text of this work; however, sufficient detail for a thorough understanding of the program functions is not provided. This appendix provides the necessary detail in the form of an example which demonstrates most of the program features. The device simulated is an N-channel Silicon MOSFET with the following parameters:

gate length	1.5 $\mu\text{m}$
gate oxide	500 $\text{\AA}$
field oxide	4000 $\text{\AA}$
junction depth	.4 $\mu\text{m}$
substrate doping	$2 \times 10^{15} \text{ cm}^{-3}$
channel implant dose	$3.4 \times 10^{11} \text{ cm}^{-2}$
fixed interface charge	$1 \times 10^{10} \text{ cm}^{-2}$
gate material	n-type polysilicon
source/drain contacts	aluminum.

Figures A.1 and A.2 show the input deck for the program. The deck is divided into two parts to demonstrate the saving and restoring of program data files. Usually, one would split the deck into several pieces so that each step in the simulation may be verified before proceeding to the next.

The first item in each line is the *card* name and the remaining items are *parameters*. There are three types of parameters—numeric, alphanumeric, and logical. Numeric parameters are followed by an equal sign and a numeric value. Alphanumeric parameters are also followed by an equal sign but may

```

TITLE      MOSFET EXAMPLE
$ *** Generate mesh ***
MESH       RECTANGULAR NX=31 NY=22 OUTFILE=MESH1 BULKDOP=2E15 P.TYPE
X.MESH     NODE=1 LOCATION=1 RATIO=1
X.M        N=3 L=1.5 R=.71
X.M        N=9 L=1.9 R=.8
X.M        N=15 L=2.25 R=1.11
X.M        N=21 L=2.6 R=.9
X.M        N=27 L=3 R=1.25
X.M        N=29 L=3.5 R=1.414
X.M        N=31 L=4.5 R=1.414
Y.M        N=1 L=-.05 R=1
Y.M        N=4 L=0 R=1
Y.M        N=22 L=3 R=1.25
$ *** Expand field oxide ***
SPREAD     LEFT WIDTH=.5 UPPER=1 LOWER=4 THICKNESS=.4 ENCROACH=1
+          VOL.RAT=.4
SPR        RIGHT W=1.5 UP=1 LO=4 THICK=.4 ENCR=1 VOL=.4
$ *** Match junctions ***
SPR        LEFT W=.8 UP=4 LO=10 Y=.41 ENCR=.9 GRADING=.7
SPR        RIGHT W=1.8 UP=4 LO=10 Y=.41 EN=.9 GR=.7
$ *** Identify insulator and semiconductor regions ***
REGION     NUMBER=1 X.LOW=1 X.HIGH=31 Y.LOW=1 Y.HIGH=4 INSULATOR
REG        NUM=2 X.L=1 X.H=31 Y.L=4 Y.H=22 SEMICONDUCTOR
$ *** Dope the semiconductor ***
$          *** substrate ***
DOPING     P.TYPE CONCENTR=2E15 UNIFORM
$          *** channel implant ***
DOP        P DOSE=3.4E11 Y.PEAK=.18 Y.CHARAC=.2404 GAUSSIAN
$          *** source and drain ***
DOP        DONOR CONC=4E19 LEFT.JUN Y.JUNC=.4 Y.PEAK=0 GAUSS
+          X.RIGHT=1.5 XY.RATIO=1
DOP        DONOR CONC=4E19 RIGHT.J Y.J=.4 Y.P=0 GAUSS X.L=3 XY=1
$ *** Fixed surface states ***
QF         CONCENTR=1E10 X.LOW=3 X.HIGH=27 Y.LOW=4 Y.HIGH=4
$ *** Identify electrode locations ***
ELECTROD   NUMBER=1 X.LOW=3 X.HIGH=27 Y.LOW=1 Y.HIGH=1
ELEC       N=2 X.L=1 X.H=31 Y.L=22 Y.H=22
ELEC       N=3 X.L=1 X.H=2 Y.L=4 Y.H=4
ELEC       N=4 X.L=28 X.H=31 Y.L=4 Y.H=4
$ *** Print vertical grid info ***
PRINT      POINTS IX.MIN=15 IX.MAX=15
$ *** Plot grid and junctions ***
FLOT.2D    X.MIN=1 X.MAX=4.5 Y.MIN=-.3 Y.MAX=3 NO.TOP BOUNDARY
+          JUNCTION GRID
$ *** End ***

```

Fig. A.1. Sample PISCES input card deck for mesh generation and device structure definition.

```

TITLE      PERFORM SOLUTIONS
$ *** Get mesh ***
MESH       IN=MESH1
$ *** Perform symbolic matrix factorization ***
SYMB       OUT=SYMB1
$ *** Prepare for initial solution ***
SETUP      INIT PRINT TEMPERAT=300 P.ELECT=2
$ *** specify materials ***
MATERIAL   NUMBER=1 OXIDE
MATER      NUM=2 SILICON
$ *** specify contacts ***
CONTAC     NUMBER=1 N.POLY
CONTAC     NUM=2 NEUTRAL
CONTAC     NUM=3 ALUMINUM
CONTAC     NUM=4 ALUM
$ *** Specify mobility models ***
MOBILITY   VSAT CONMOB
$ *** Solve initial solution ***
SOLV       PRINT OUT=EXOUT0
$ *** Step gate bias ***
SETUP      INF=EXOUT0 PREVIOUS V1=0
SOLVE      PRINT
SETUP      PROJECT V1=2
SOLVE      OUTFILE=EXOUT1 PRINT
$ *** Step drain bias ***
SETUP      PREVIOUS V4=.5
SOLVE      SINGLE ACCEL OUT=EXOUT2 PRINT
SETUP      PROJECT VSTEP=.5 NSTEPS=3 ELEC=4
SOLVE      SINGLE ACCEL OUT=EXOUT3 PRINT
$ *** Plot results at Vg=2, Vd=2 ***
PLOT.2     X.MIN=1 X.MAX=4.5 Y.MIN=-.3 Y.MAX=3 BOUND NO.TOP JUNC DEPL
CONTOUR     POTENTIAL MIN.VAL=-.2 MAX.VAL=1.6 DEL.VAL=.2
PLOT.2     X.MIN=1 X.MAX=4.5 Y.MIN=-.3 Y.MAX=3 BOUND NO.TOP JUNC DEPL
CONTO      QF.POT MIN=.2 MAX=2 DEL=.2
$ *** End ***

```

Fig. A.2. Sample PISCES input card deck for specifying device material characteristics and obtaining simulation solutions.

have any alphanumeric character as a value. Logical parameters may be followed by an equal sign and the words *true* or *false* or may appear alone in which case they are assigned a logical value of *true*. A "+" in the first column indicates a continuation of the previous line. Note that either card names or parameters names may be shortened if the resulting name is unambiguous.

The card and parameter names recognized by the program (8 characters maximum) are shown in upper case letters with the remainder of the name in lower case letters for clarity. The remainder of this appendix details the use of cards and parameters by describing their use in the sample input decks of Figures A.1 and A.2.

### TITLE

The TITLE card has no parameters. All of the characters after the card name are stored and used as a header for all printed listings.

### \$ or COMMENT

Either \$ or COMMENT may be used to specify a comment line which is ignored by the program.

### MESH

The MESH card indicates the beginning of a sequence of cards serving to define the device structure. The sequence is terminated when a non-mesh-defining card is encountered. Most of the cards must appear in the order given in order to properly define the device. A RECTANGULAR mesh (grid) is specified meaning that the grid nodes will initially lie at the intersections of parallel horizontal and vertical lines. Distortion of this grid is allowed later. There are 31 vertical grid lines (NX) and 22 horizontal grid lines (NY). At the termination of the mesh sequence, all of the structure data will be stored in a file (OUTFILE) called MESH1. The substrate doping (BULKDOP) is  $2 \times 10^{15} \text{ cm}^{-3}$  and is p-type (P.TYPE). The substrate may also be specified as N.TYPE. If a mesh file has been previously stored, all structure data may be read with the single parameter INFILE and the name of the file.

## X.MESH

The X.MESH card specifies the location along the  $x$  axis of one of the vertical grid lines. The first NODE (actually all nodes on the first vertical line) has an  $x$ -axis coordinate (LOCATION) of  $1\text{ }\mu\text{m}$ . The origin of the  $x$  axis may be arbitrarily chosen. The RATIO value has no meaning for the first node.

The next card is also an X.MESH card but the card name and parameter names are shortened for ease of typing. All of the cards in this example follow this same pattern in which full card and parameter names are used on the first occurrence of a card type, but shortened names are used thereafter. The second X.MESH card places node 3 at  $1.5\text{ }\mu\text{m}$ . The RATIO parameter specifies that the spacing between vertical grid lines 2 and 3 should be only .71 of the spacing between lines 1 and 2. If there were more grid lines specified in the interval, then each successive space (from left to right) would be .71 as large as the previous space. The additional X.MESH cards specify the remaining grid lines, locations and spacing ratios up to the rightmost edge of the simulation region at  $4.5\text{ }\mu\text{m}$ .

## Y.MESH

The Y.MESH card serves the same function as the X.MESH card but in the orthogonal direction. The  $y$  axis is positive downward. The first horizontal grid line is placed at the top of the gate oxide. The fourth grid line is placed at the oxide-semiconductor interface and is chosen as the origin. The last grid line is placed at the bottom of the simulation region,  $3\text{ }\mu\text{m}$  deep into the substrate.

## SPREAD

The SPREAD card is used to distort the grid in the vertical direction in order to match device surface or interface shapes or other internal device structure. The operation results in horizontal grid lines which are vertically displaced on either the left or right side of the device with a smooth variation of this displacement across the device. The first two spread cards expand the oxide region on the LEFT and RIGHT sides of the device from the gate oxide thickness to the field oxide THICKNESs of 4000 Å. The second two distort the grid near the semiconductor surface to match the source/drain junction profiles. The first spread card expands the grid between lines 1 (UPPER) and 4 (LOWER) specified later as the oxide region, for a WIDTH of .5  $\mu\text{m}$  from the LEFT edge of the device. The ENCROACHment factor specifies the abruptness of the transition from spread to non-spread grid. The VOL.RATio parameter specifies the ratio of the downward displacement of the lower grid line to the net increase in thickness, corresponding to the volume ratio of silicon consumed to oxide grown in thermal oxidation.

In the third and fourth SPREAD cards, the UPPER grid line, the oxide-semiconductor interface, is not moved but the LOWER line is moved to the Y.LOWER coordinate of .41  $\mu\text{m}$  which is just below the source/drain junctions. The spacing of all grid lines in between is changed to a GRADING of .7 in order to provide the proper grid placement on the steep source/drain impurity profiles. The GRADING value is used exactly like the RATIO parameter in the X.MESH card.

## REGION

The REGION card is used to define the INSULATOR and SEMICONDUCTor regions of the device. It can also be used to rigidly restrict the region of



the device which receives impurity doping. Each region must be sequentially numbered using the NUMBER parameter. The X.LOW, X.HIGH, Y.LOW, and Y.HIGH parameters specify the grid lines which bound the region. In this typical example, the entire semiconductor substrate is contained in one region, and the oxide in another.

### DOPING

The DOPING card is used to add impurities to the device within the bounds of the most recent REGION card. The first DOPING card specifies the substrate doping to be a UNIFORM distribution of P.TYPE impurity with a CONCENTRATION of  $2 \times 10^{15} \text{ cm}^{-3}$ .

The second DOPING card specifies the channel implant as a GAUSSIAN implant of a P.TYPE impurity with a DOSE of  $3.4 \times 10^{11} \text{ cm}^{-2}$ , a peak (Y.PEAK) at  $.18 \mu\text{m}$  and CHARACTERISTIC length ( $\sqrt{2}\sigma$ ) of  $.2401 \mu\text{m}$ .

The third DOPING card specifies a GAUSSIAN source doping with a peak CONCENTRATION of  $4 \times 10^{19} \text{ cm}^{-3}$  DONOR impurities, with the peak (Y.PEAK) at the origin and the junction (Y.JUNCTI) at  $.4 \mu\text{m}$  computed using the background doping on the left side (LEFT.JUN) of the device. The impurity distribution is specified to be uniform in the lateral direction from the left edge of the device (by default) to the  $1.5 \mu\text{m}$  location on the  $x$ -axis (X.RIGHT). This point corresponds to the right edge of a diffusion or implant window. The lateral profile beyond the  $1.5 \mu\text{m}$  coordinate is also Gaussian (by default) and the characteristic length in the  $x$  direction is specified by XY.RATIO to be equal to the characteristic length in the  $y$  direction, resulting in cylindrical junctions.

The fourth DOPING card specifies the same doping profile for the drain. Additional parameters allow the impurity type to be specified as N.TYPE or

ACCEPTOR and permit a complementary error function lateral impurity profile by specifying X.ERFC.

### QF

The QF card is used to specify the fixed surface states charge at an insulator-semiconductor interface. The CONCENTRATION is  $1 \times 10^{10} \text{ cm}^{-2}$  and exists along horizontal grid line number 4 (Y.LOW, Y.HIGH) from vertical grid line number 3 (X.LOW) to vertical grid line number 27 (X.HIGH).

### ELECTRODe

The ELECTRODe card specifies nodes in the grid at which the potential boundary conditions will be applied. Each group of nodes is assigned a NUMBER which is used to reference the group. The node clusters are specified by the bounding grid lines as in the QF card (X.LOW, X.HIGH, Y.LOW, Y.HIGH). Electrode number 1 is the gate, 2 is the substrate, 3 is the source and 4 is the drain.

This is the last mesh-defining card. Reading the next card terminates the MESH sequence and causes a final computation of mesh parameters and stores the mesh data in the specified output file.

### PRINT

The PRINT card provides terminal or line printer listings of a large variety of simulation information. The information is printed only for areas of the device within a window. The window may be specified by providing the grid line boundaries (IX.MIN, IX.MAX, IY.MIN, IY.MAX) or coordinate boundaries (X.MIN, X.MAX, Y.MIN, Y.MAX). The POINTS parameter prints data associated with nodes in the grid. Other allow-

able PRINT parameters are: ELEMENTS—nodes composing each triangle; GEOMETRY—coupling coefficients; SOLUTION—potential, quasi-FERMI potential and electron concentration at each node; MATERIAL—detail on the material parameters for each device region.

### PLOT.2D

The PLOT.2D card makes a two-dimensional plot of specified device characteristics. The plot window is specified by coordinates (X.MIN, X.MAX, Y.MIN, Y.MAX), which may lie inside or outside of the simulation region. The BOUNDARY parameter plots the device external boundary, interfaces and electrodes, JUNCTION plots the metallurgical junctions, GRID plots the triangular grid, and DEPL.EDGE plots the depletion edges. NO.TOP inhibits the plotting of tic marks across the top of the plot and NO.TIC inhibits all tic marks. NO.CLEAR inhibits clearing of the display between plots to allow superimposed plots. A plot file may be generated by specifying OUTFILE and a file name. The resulting plot is shown in Figure A.3. Further plotting capability is provided by the CONTOUR card to be described later.

Figure A.2 shows the card sequence which performs the solutions on the device. In this card sequence, the MESII card merely reads the file MESII which contains all of the necessary device structure information.

### SYMBOLIC

The SYMBOLIC card invokes the symbolic factorization of the coefficient matrix for the LU decomposition. INFILE and OUTFILE are used to read or write the pointer arrays resulting from the factorization. The DISSECT parameter may be used to perform a nested dissection renumbering on the

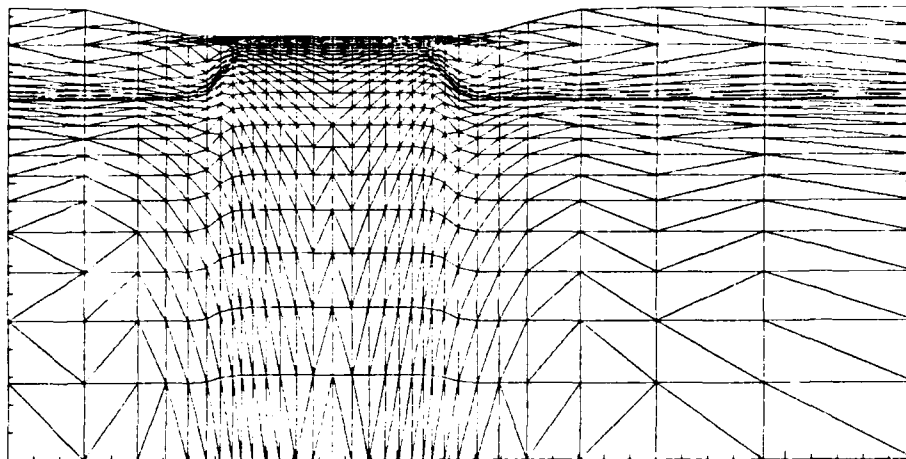


Fig. A.3. Plot of the mesh generated by the PISCES example.

grid. Alternatively, MINIMIZE may be used to renumber the grid by rows if there are fewer nodes across a row than down a column. The PRINT parameter prints a summary of relevant factorization parameters.

### SETUP

The SETUP card computes the coefficient matrix and the initial guess prior to every solution. It may be followed by a sequence of cards which specify or modify various device parameters. The sequence is terminated when a non-setup-sequence card is encountered. Since this is the first solution performed on this device the INITIAL parameter is specified. This results in a charge-neutral initial guess and a flat-band bias assignment. The PRINT parameter invokes a listing of the SETUP parameter values on termination of the SETUP sequence. The device TEMPERATURE is specified to be  $300^{\circ}K$ . The P.ELECTROde parameter assigns the substrate bias (electrode number 2) as the hole quasi-Fermi level value. An alternative way to set the hole

quasi-Fermi level is to explicitly specify it with the P.BIAS parameter.

In the second occurrence of the SETUP card (several cards down the list), the PREVIOUS solution is used as the initial guess for the next solution. The previous solution is read using INFILE. Actually, reading the solution from the file was not necessary since the two most recent solutions are always stored in memory. INFILE reads a stored solution into the most-recent-solution array and IN2FILE reads into the second-most-recent solution array. The device electrode bias levels are set using V1 through V9 and VTEN where the number corresponds to the electrode number. Here the gate is set to zero volts. All electrode voltages not explicitly set are kept at their previous values.

The third occurrence of the SETUP card demonstrates the use of the PROJECTION parameter for extrapolating an initial guess from two previous solutions. Only one electrode bias is allowed to change between the new and two previous solutions. Here the gate (electrode 1) has previous values of flat-band and zero volts and a new value of two volts.

The fourth SETUP card sets the drain voltage to .5 volts.

The fifth SETUP card demonstrates the bias stepping capability of PISCES. VSTEP sets the bias step size, NSTEPS sets the number of bias steps, and ELECTRODe specifies the electrode number (the drain) being varied. When bias steps are specified in this way, only one SETUP/SOLVE combination is required for the range of bias steps requested.

## MATERIAL

The MATERIAL card is used to specify the materials and physical parameters to be used for the simulation. Material NUMBER 1 corresponding to region number 1 is OXIDE. Other insulator specifications allowed are SiO<sub>2</sub>,

NITRIDE, SI3N4, SAPPHIRE, and INSULATOR. The relative permittivity is appropriately set for each of the insulator parameters except INSULATOR which requires an explicit PERMITTIVITY value. It is optional for the other specifications.

The second MATERIAL card assigns SILICON physical parameters to region NUMBER 2. Other semiconductor specifications allowed are gallium arsenide (GAAS) and SEMICONDUCTOR. A variety of physical parameters are set by the SILICON or GAAS parameters or may be optionally set, but they must be explicitly set for the SEMICONDUCTOR parameter. NI300 and EG300 are the intrinsic carrier concentration and energy gap at 300°K, the PERMITTIVITY and electron AFFINITY may be specified, and TP and TN are the hole and electron minority carrier lifetimes. A constant MOBILITY and a carrier saturation velocity (VSAT) may also be specified. EGALPHA and EGBETA are terms in the expression for energy gap variations with temperature:

$$E_g(T) = E_g(300) + \alpha \left( \frac{1}{1 + \beta} - \frac{(T/300)^2}{T/300 + \beta} \right) \quad (4.01)$$

where  $E_g$  is the energy gap,  $T$  is the temperature in °K,  $\alpha$  is EGALPHA, and  $\beta$  is EGBETA. These values are related to those of Sze [A.1] by  $\alpha = 300\alpha(\text{Sze})$  and  $\beta = \beta(\text{Sze})/300$ .

## CONTACT

The CONTACT card specifies the type of material used for the device electrodes. The NUMBER corresponds to the ELECTRODE number. Alternatively, ALL may be used to specify with one card that all of the contacts use the same material. In the example, the gate is N.POLYSILICON and the source and drain are ALUMINUM. Other allowed materials

are P.POLYSilicon, MOLYBDENum, and TUNGSTEN, molybdenum disilicide (MO.DISIL) and tungsten disilicide (TU.DISIL). Alternatively, the WORKFUNction may be provided explicitly. It is very useful in those situations where the contact characteristics do not influence device operation to specify a NEUTRAL contact. This specification guarantees that there will be no carrier accumulation or depletion at the contact. The substrate contact in the MOSFET simulation, for example, is specified in this manner since the simulated substrate contact at the bottom of the simulation region is much closer to the surface than the actual substrate contact.

### MOBILITY

The MOBILITY card is used to specify which of the carrier transport and recombination models are to be used in the simulation. This capability is used primarily to aid comparison of PISCES to other device simulation programs which do not have the models. For normal simulations one would turn on all of the models. The example specifies that velocity saturation (VSATURAT) and impurity concentration dependent mobility (CONMOB) be used in the simulation. The other model parameter allowed is SRHIRECOM for Shockley-Read-Hall recombination. Each of these specifications remains in force until terminated with a NOVSATUR, NOCONMOB, or NOSRHIREC specification.

The MOBILITY card is the last card in the SETUP sequence. Reading of the next card causes the initial guess to be generated and all parameters to be set as specified.

### SOLVE

The SOLVE card controls the method of solution used in the simulation. The first SOLVE card in the example specifies that iteration information (bias,

charge and current at each electrode) be PRINTed and that the solution be saved in a file called EXOUT0 (OUTFILE). The PRINT parameter can be terminated using NOPRINT.

The fourth SOLVE card specifies that the SINGLEPoisson iteration method be used since significant drain bias is being applied. The default condition is the MULTIPoisson iteration method. One may also specify POISSON only iterations in which no continuity equation solutions are performed. The default condition is BOTH where both sets of equations are solved. The ACCELERation parameter specifies that the linearization factor reduction method of convergence acceleration discussed in the text is to be used. The default is NOACCElation. The alternate acceleration method, overRELAXation is also available with the default of NORELAX. There are four levels of convergence available, COARSE, MEDIUM, FINE and LIMIT with MEDIUM being the default. Each of the first three have succeedinglly tighter convergence limits. The fourth level, LIMIT, specifies that a number of iterations equal to ITLIMIT be executed regardless of the level of convergence. Alternatively, the actual iteration convergence tolerances themselves P.TOLERance and C.TOLERance may be specified.

The fifth SOLVE card demonstrates its use in the bias stepping mode. No special consideration is required; however, the file name specified by the OUTFILE parameter will be incremented by one character/digit for each bias step. Thus 3 different solution files will be saved for this bias stepping sequence; EXOUT3, EXOUT4 and EXOUT5.

The PLOT.2D card has been covered earlier with the card sequence of Figure A.1.



## CONTOUR

The CONTOUR card is used to plot two-dimensional contours of various device parameters. EquiPOTENTIAL contours are plotted at potential values of MIN.VALUE to MAX.VALUE with DEL.VALUE steps. Other device contours which may be plotted are quasi-Fermi potential (QF.POTEN), DOPING, ELECTRON concentration, HOLE concentration, net charge concentration (NET.CHRG) and net carrier concentration (NET.CARR). Logarithmic contour intervals may be specified by LOGARITHm. The values for MIN.VALUE, MAX.VALUE, and DEL.VALUE are then the logarithms of the desired values. In order to plot logarithmic intervals of negative values of NET.CHRG or NET.CARR, the NEGATIVE parameter must be specified. The line type to be used in the contour plot may be specified using the LINE.TYPe parameter and an integer value between one and 11. Figures A.4 and A.5 show the potential and quasi-Fermi potential contour plots generated by the CONTOUR cards.

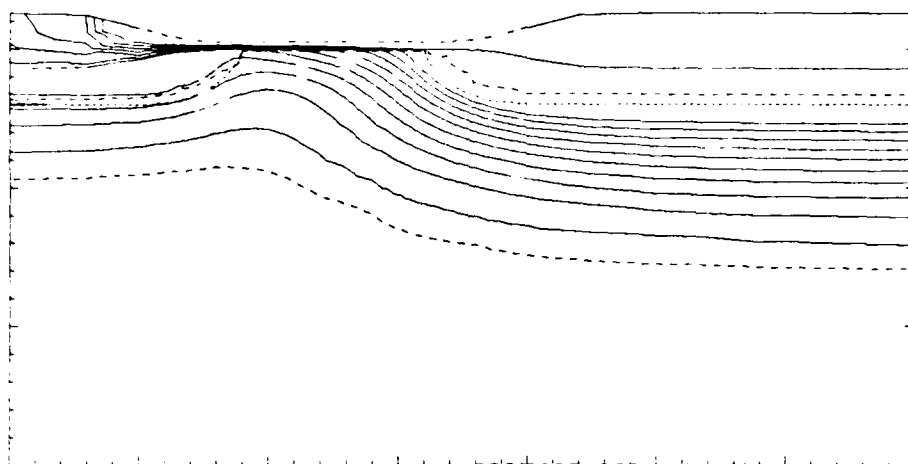


Fig. A.4. Equipotential contours generated by the example at  $V_G = 2V$  and  $V_{DS} = 2V$ .

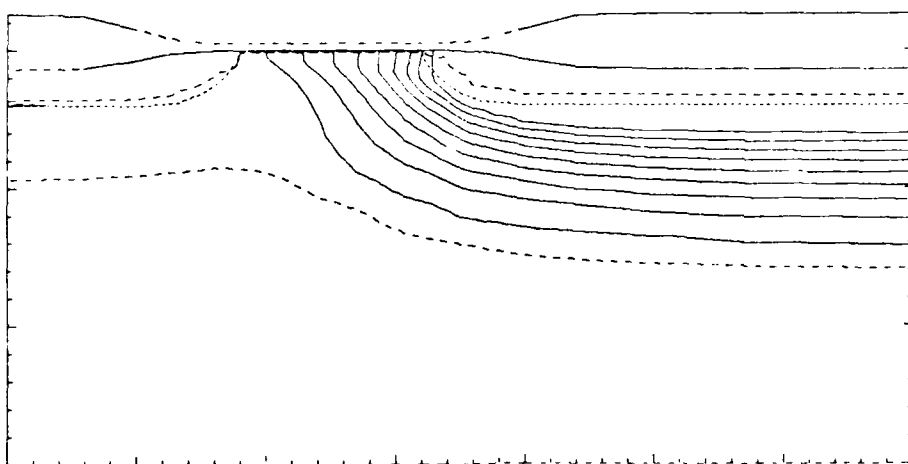


Fig. A.5. Quasi-Fermi potential contours generated by the example at  $V_G = 2V$  and  $V_{DS} = 2V$ .

## REFERENCES

- [1.1] R. W. Keyes, "The Evolution of Digital Electronics Towards VLSI," *IEEE Trans. on Electron Devices*, ED-26, 1979, p.271.
- [1.2] J. G. Posa, "VIISIC Proposals Take Six Fast Tracks," *Electronics*, 54, No. 19, 1981, p.89.
- [1.3] R. H. Dennard, *et al.*, "Design of Ion-Implanted MOSFET's with Very Small Physical Dimensions," *IEEE J. Solid-State Circuits*, SC-9, 1974, p.256.
- [1.4] L. D. Yau, "A Simple Theory to Predict the Threshold Voltage of Short-Channel IGFET's," *Solid-State Electronics*, 17, 1974, p.1059.
- [1.5] G. W. Taylor, "Subthreshold Conduction in MOSFET's," *IEEE Trans. on Electron Devices*, ED-25, 1978, p.337.
- [1.6] J. Nishizawa, T. Terasaki, and J. Shibata, "Field Effect Transistor versus Analog Transistor (Static Induction Transistor)," *IEEE Trans. on Electron Devices*, ED-22, 1975, p.185.
- [1.7] P. K. Chatterjee, G. W. Taylor and M. Malwah, "Taper Isolated Dynamic Grain RAM cell," *Technical Digest, Intl. Electron Devices Mtg.*, 1978.
- [1.8] S. P. Fan, *et al.* "MOTIS-C: A New Circuit Simulator for MOS LSI Circuits," *Proc. IEEE Intl. Symp. Circuits and Systems*, 1977, p.700.
- [1.9] D. A. Antoniadis, *et al.*, SUPREM 1--A Program for IC Process Modeling and Simulation, TR No. 5019-1, Stanford University, Stanford, CA, 1977.

- [1.10] L. W. Nagel, "SPICE2, A Computer Program to Simulate Semiconductor Circuits," University of California, Berkeley, ERL Memo ERL-M520, 1975.
- [1.11] H. G. Lee, Two-Dimensional Impurity Diffusion Studies: Process Models and Test Structures for Low-Concentration Boron Diffusion, TR No. G201-8, Stanford Electronics Laboratories, Stanford University, Stanford, CA, 1980.
- [1.12] H. G. Lee and R. W. Dutton, "Measurement of Two-Dimensional Profiles Near Locally Oxidized Regions," *Tech. Digest, Intl. Electron Devices Meeting*, 1975, p.65.
- [1.13] J. J. Barnes, K. Shimohigashi, and R. W. Dutton, "Short Channel MOSFET's in the Punchthrough Current Mode," *IEEE Trans. on Electron Devices*, ED-26, 1979, p.446.
- [1.14] H. K. Gummel, "A Self-Consistent Iterative Scheme for One-Dimensional Steady State Transistor Calculations," *IEEE Trans. on Electron Devices*, ED-11, 1964, p.455.
- [1.15] G. Dahlquist, Å. Björck, and N. Anderson, *Numerical Methods*, Prentice-Hall, Englewood Cliffs, NJ, 1974, p.224.
- [1.16] A. de Mari, "An Accurate Numerical Steady-State One-Dimensional Solution of the p-n Junction," *Solid-State Electronics*, 11, 1968, p.33.
- [1.17] A. de Mari, "An Accurate Numerical One-Dimensional Solution of the p-n Junction under Arbitrary Transient Conditions," *Solid-State Electronics*, 11, 1968, p.1021.

- [1.18] D. L. Scharfetter and H. K. Gummel, "Large-Signal Analysis of a Silicon Read Diode Oscillator," *IEEE Trans. on Electron Devices*, 16, 1969, p.64.
- [1.19] G. E. Possin, M. S. Adler, and B. J. Baliga, "Measurement of Heavy Doping Parameters in Silicon by Electron-Beam-Induced Current," *IEEE Trans. on Electron Devices*, ED-27, 1980, p.983.
- [1.20] F. F. Fang and A. B. Fowler, "Transport Properties of Electrons in Inverted Silicon Surfaces," *Physical Review*, 169, 1968, p.620.
- [1.21] B. V. Gokhale, "Numerical Solutions for a One-Dimensional Silicon n-p-n Transistor," *IEEE Trans. on Electron Devices*, ED-17, 1970, p.594.
- [1.22] G. D. Hachtel, R. C. Joy, and J. W. Cooley, "A New Efficient One-Dimensional Analysis Program for Junction Device Modelling," *Proc. of the IEEE*, 60, 1972, p.86.
- [1.23] D. C. D'Avanzo, M. Vanzi, and R. W. Dutton, One-Dimensional Semiconductor Device Analysis (SEDAN), TR No. G-201-5, Stanford Electronics Laboratories, Stanford, CA, 1979.
- [1.24] D. P. Kennedy and R. R. O'Brien, "Two-Dimensional Mathematical Analysis of a Planar Type Junction Field Effect Transistor," *IBM J. Res. Develop.*, 13, 1969, p.662.
- [1.25] D. P. Kennedy and R. R. O'Brien, "Computer-Aided Two Dimensional Analysis of the Junction Field-Effect Transistor," *IBM J. Res. Develop.*, 14, 1970, p.95.

- [1.26] D. P. Kennedy and R. R. O'Brien, "Two-Dimensional Analysis of J.F.E.T. Structures Containing a Low-Conductivity Substrate, " *Electronics Letters*, 7, 1971, p.714.
- [1.27] J. W. Slotboom, "Iterative Scheme for 1- and 2-Dimensional DC Transistor Simulation," *Electronics Letters*, 5, 1969, p.677.
- [1.28] J. W. Slotboom, "Computer-Aided Two-Dimensional Analysis of Bipolar Transistors," *IEEE Trans. on Electron Devices*, ED-20, 1973, p.699.
- [1.29] M. Reiser, "Two Dimensional Analysis of Substrate Effects in Junction F.E.T.'s," *Electronics Letters*, 6, 1970, p.493.
- [1.30] M. Reiser, "Difference Methods for the Solution of the Time-Dependent Semiconductor Flow Equations," *Electronics Letters*, 7, 1971, p.353.
- [1.31] M. Reiser and P. Wolf, "Computer Study of Submicrometre F.E.T.'s," *Electronics Letters*, 8, 1972, p.254.
- [1.32] M. Reiser, "A Two-Dimensional Numerical FET Model for DC, AC, and Large Signal Analysis," *IEEE Trans. on Electron Devices*, ED-20, 1973, p.35.
- [1.33] M. Reiser, "On the Stability of Finite Difference Schemes in Transient Semiconductor Problems," *Computer Meth. in Appl. Mech. and Eng.*, 2, 1973, p.65.
- [1.34] M. Reiser, "Computing Methods in Semiconductor Problems," IBM Report RJ 1343 (20931), 1974.

- [1.35] G. D. Hachtel and M. H. Mack, "A Graphical Study of the Current Distribution in Short-Channel IGFET's," *Digest of Technical Papers*, Intl. Solid-State Circuits Conf., 1973, p.110.
- [1.36] G. D. Hachtel, M. H. Mack, and R. R. O'Brien, "Semiconductor Device Analysis via Finite Elements," Eighth Asilomar Conf. on Circuits and Systems, 1974.
- [1.37] J. J. Barnes and R. J. Lomax, "Two-Dimensional Finite-Element Simulation of Semiconductor Devices," *Electronics Letters*, 10, 1974, p.341.
- [1.38] J. J. Barnes and R. J. Lomax, "Finite-Element Methods in Semiconductor Device Simulation," *IEEE Trans. on Electron Devices*, ED-24, 1977, p.1082.
- [1.39] E. M. Buturla and P. E. Cottrell, "Two-Dimensional Finite- Element Analysis of Semiconductor Steady State Transport," Intl. Conf. on Computational Methods in Non-Linear Mechanics, 1974.
- [1.40] P. E. Cottrell and E. M. Buturla, "Steady State Analysis of Field Effect Transistors via the Finite-Element Method," *Technical Digest*, Intl. Electron Devices Meeting, 1975, p.51.
- [1.41] P. E. Cottrell and E. M. Buturla, "Two-Dimensional Static and Transient Simulation of Mobile Carrier Transport in a Semiconductor," *Proceedings of the NASECODE I Conf.*, 1979, p.31.
- [1.42] J. A. Greenfield and R. W. Dutton, "Non-planar VLSI Device Analysis Using the Solution of Poisson's Equation," *IEEE J. Solid-State Circuits*, SC-15, 1980, p.585.

- [1.43] R. W. Dutton and S. E. Hansen, "Process Modeling of Integrated Circuit Device Technology," *Proceedings of the IEEE*, 69, 1981, p.1305.
- [1.44] M. S. Mock, "A Two-Dimensional Mathematical Model of the Insulated-Gate Field-Effect Transistor," *Solid State Electronics*, 16, 1973, p.601.
- [1.45] S. Selberherr, W. Fichtner, and W. H. Pötzl, "MINIMOS- A Program Package to Facilitate MOS Device Design and Analysis," *Proc. of the NASECODE I Conf.*, 1979.
- [1.46] S. Liu, B. Hoefflinger, and D. O. Pederson, "Interactive Two-Dimensional Design of Barrier Controlled MOS Transistors," *IEEE J. Solid-State Circuits*, SC-15, 1980, p.615.
- [1.47] E. M. Buteria, *et al.*, "Three-Dimensional Finite Element Simulation of Semiconductor Devices," *Digest of Technical Papers*, Intl. Solid State Circuits Conf., 1980, p.76.
- [1.48] A. Yoshii, S. Horiguchi, and T. Sudo, "A Numerical Analysis for Very Small Semiconductor Devices," *Digest of Technical Papers*, Intl. Solid-State Circuits Conf., 1980, p.80.
- [1.49] S. G. Chamberlain and A. Husain, "Three-Dimensional Simulation of VLSI MOSFET's: The Three-Dimensional Simulation Program WATMOS," *Technical Digest*, Intl. Electron Devices Meeting, 1981, p.592.



- [1.50] P. E. Cottrell and E. M. Buturla, "Steady State Analysis of Field Effect Transistors via the Finite-Element Method," *Technical Digest, Intl. Electron Devices Meeting, 1975*, p.51.
- [1.51] S. Y. Oh and R. W. Dutton, "A Simplified Two-Dimensional Analysis of MOS Devices," *Proc. of the NASECODE I Conf., 1979*.
- [1.52] A. M. Winslow, "Numerical Solution of the Quasilinear Poisson Equation in a Nonuniform Triangle Mesh," *J. Computational Physics*, 2, 1967, p.149.
- [2.1] M. S. Adler, "Factors Determining Forward Voltage Drop in the Field-Terminated Diode (FTD), *IEEE Trans. on Electron Devices, ED-25*, 1978, p.529.
- [2.2] A. D. Sutherland, A Two-Dimensional Computer Model for the Steady-State Operation of MOSFET's, ECOM-75-1344-F, Supplement, 1977.
- [2.3] D. P. Kennedy and R. R. O'Brien, "Two-Dimensional Mathematical Analysis of a Planar Type Junction Field Effect Transistor," *IBM J. Res. Develop.*, 13, 1969, p.662.
- [3.1] J. A. Greenfield, R. W. Dutton, "Nonplanar VLSI Device Analysis Using the Solution of Poisson's Equation, " *IEEE J. Solid-State Circuits, SC-15*, 1980, p.505.
- [3.2] A. M. Winslow, "Numerical Solution of the Quasilinear Poisson Equation in a Nonuniform Triangle Mesh," *J. Computational Physics*, 2, 1967, p.149.

- [3.3] D. L. Scharfetter and H. K. Gummel, "Large-Signal Analysis of a Silicon Read Diode Oscillator," *IEEE Trans. on Electron Devices*, 16, 1969, p.64.
- [4.1] D. J. Rose and R. A. Willoughby, *Sparse Matrices and Their Applications*, Plenum Press, New York, 1972.
- [4.2] G. E. Forsythe and C. B. Moler, *Computer Solution of Linear Algebraic Systems*, Prentice-Hall, Englewood Cliffs, NJ, 1967, p.27.
- [4.3] R. W. Hockney, *Methods in Computational Physics*, 9, Academic Press, New York and London, 1970.
- [4.4] T. Wada and J. Frey, "Physical Basis of Short-Channel MESFET Operation," *IEEE Trans. on Electron Devices*, ED-26, 1979, p.476.
- [4.5] R. S. Varga, *Matrix Iterative Analysis*, Prentice-Hall, Englewood Cliffs, NJ, 1962, p.58.
- [4.6] G. Dahlquist, Å. Björck, and N. Anderson, *Numerical Methods*, Prentice-Hall, Englewood Cliffs, NJ, 1974, p.188.
- [4.7] G. E. Forsythe and C. B. Moler, *Computer Solution of Linear Algebraic Systems*, Prentice-Hall, Englewood Cliffs, NJ, 1967.
- [4.8] S. Liu, B. Hoeflinger, and D. O. Pederson, "Interactive Two-Dimensional Design of Barrier Controlled MOS Transistors," *IEEE J. Solid-State Circuits*, SC-15, 1980, p.615.
- [4.9] J. A. Greenfield and R. W. Dutton, "Non-planar VLSI Device Analysis Using the Solution of Poisson's Equation," *IEEE J. Solid-State Circuits*, SC-15, 1980, p.585.

- [4.10] H. L. Stone, "Iterative Solution of Implicit Approximations of Multidimensional Partial Differential Equations," *SIAM J. on Numerical Analysis*, 5, 1968, p.530.
- [4.11] M. S. Mock, "A Two-Dimensional Mathematical Model of the Insulated-Gate Field-Effect Transistor," *Solid State Electronics*, 16, 1973, p.601.
- [4.12] D. S. Kershaw, "The Incomplete Cholesky—Conjugate Gradient Method for the Iterative Solution of Systems of Linear Equations," *J. Computational Physics*, 26, 1978, p.43.
- [4.13] P. Concus, G. H. Golub, and D. P. O'Leary, A Generalized Conjugate Gradient Method for the Numerical Solution of Elliptic Partial Differential Equations, Lawrence Berkeley Laboratory Pub. LBL-4604, Berkeley, CA, 1975.
- [4.14] J. A. George, "Nested Dissection of a Regular Finite Element Mesh," *SIAM J. on Numerical Analysis*, 10, 1973, p.585.
- [4.15] I. S. Duff, A. M. Erisman, and J. K. Reid, "On George's Nested Dissection Algorithm," *SIAM J. on Numerical Analysis*, 13, 1976, p.686.
- [4.16] D. A. Calavhan and P. G. Buning, Vectorized General Sparsity Algorithms with Backing Store, SEL Report 96, Systems Engineering Laboratory, Ann Arbor, MI, 1977.
- [4.17] E. M. Buturla and P. E. Cottrell, "Simulation of Semiconductor Transport Using Coupled and Decoupled Solution Techniques," *Solid State Electronics*, 23, 1980, p.331.

- [4.18] H. K. Gummel, "A Self-Consistent Iterative Scheme for One-Dimensional Steady State Transistor Calculations," *IEEE Trans. on Electron Devices*, ED-11, 1964, p.455.
- [4.19] B. A. Carré, "The Determination of the Optimum Accelerating Factor for Successive Over-relaxation," *Computer J.*, 4, 1961, p.73.
- [5.1] J. A. Greenfield and R. W. Dutton, "Non-planar VLSI Device Analysis Using the Solution of Poisson's Equation," *IEEE J. Solid-State Circuits*, SC-15, 1980, p.585.
- [5.2] M. S. Mock, "A Two-Dimensional Mathematical Model of the Insulated-Gate Field-Effect Transistor," *Solid State Electronics*, 16, 1973, p.601.
- [5.3] F. F. Fang and A. B. Fowler, "Transport Properties of Electrons in Inverted Silicon Surfaces," *Physical Review*, 169, 1968, p.620.
- [5.4] A. Hartstein, T. H. Ning, and A. B. Fowler, "Electron Scattering in Silicon Inversion Layers by Oxide and Surface Roughness," *Surface Science*, 58, 1976, p.178.
- [5.5] A. G. Sabnis and J. T. Clemens, "Characterization of the Electron Mobility in the Inverted  $\langle 100 \rangle$  Si Surface," *Technical Digest, Intl. Electron Devices Meeting*, 1979, p.18.
- [5.6] R. S. Muller and T. I. Kamins, *Device Electronics for Integrated Circuits*, John Wiley and Sons, New York, NY, 1977, p.376.

- [5.7] A. P. Gnädinger and H. E. Tally, "Quantum Mechanical Calculation of the Carrier Distribution and the Thickness of the Inversion Layer of the MOS Field-Effect Transistors," *Solid State Electronics*, 13, 1970, p.1301.
- [5.8] S. C. Sun, "Electron Mobility in Inversion and Accumulation Layers on Thermally Oxidized Silicon Surfaces," *IEEE Trans. on Electron Devices*, ED-27, 1980, p.1497.
- [A.1] S. M. Sze, *Physics of Semiconductor Devices*, Wiley-Interscience, New York, NY, 1969, p.24.

LEED 8